PATENT ABSTRACTS OF JAPAN

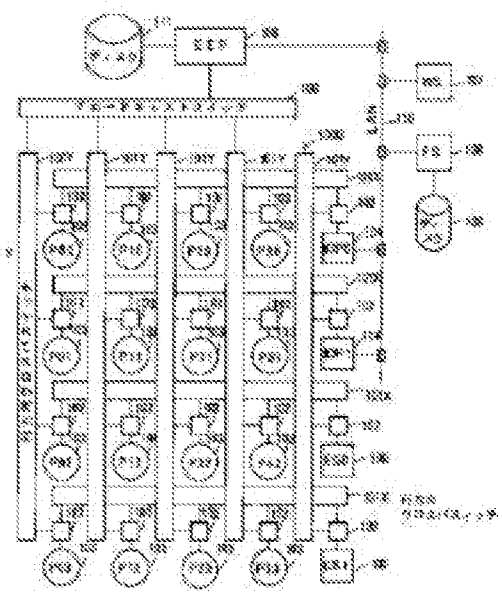
(11)Publication number : **05-081216**
(43)Date of publication of application : **02.04.1993**

(51)Int.Cl. G06F 15/16
G06F 15/16

(21)Application number : **03-241092**
(22)Date of filing : **20.09.1991**

(71)Applicant : **HITACHI LTD**
HITACHI VLSI ENG CORP
(72)Inventor : **NAKAKOSHI JUNJI**
HAMANAKA NAOKI
CHIBA HIROYUKI
HIGUCHI TATSUO
SHUDO SHINICHI
OGATA YASUHIRO
TAKEUCHI SHIGEO
TOBA TATSU

(54) **PARALLEL PROCESSOR**



(57)Abstract:
PURPOSE: To provide a parallel processor capable of attaining a parallel processing managing function and an I/O function without reducing the number of processors for executing parallel processing and the loading method of the parallel processor.
CONSTITUTION: Cross bar switches 101 having 2n+1 ports are arrayed and a group of 2n processors 103 is connected to the cross bar switches 101. Processors 104 to 106 for executing the parallel processing managing function and the I/O function are connected to the residual switches 101 out of respective switches 101 and respective processors are connected to the switches 101 through cross-over switches 102. Consequently parallel processing is executed by the 2n+1 processors and the parallel processing managing function and the I/O function are executed by other processors independently of the parallel processing. In the case of loading the parallel processor, each processor in a processor group connected to a cross bar switch in a certain dimension is successively loaded.

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平5－81216

(43)公開日　平成５年(1993)４月２日

(51)Int.Cl.[5]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ　15/16	４００　Ｙ	9190－5L		
	３９０　Ｔ	9190－5L		

審査請求　未請求　請求項の数15(全 28 頁)

(21)出願番号　特願平3－241092

(22)出願日　平成３年(1991)９月20日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(71)出願人　000233468
日立超エル・エス・アイ・エンジニアリング株式会社
東京都小平市上水本町５丁目20番１号

(72)発明者　中越　順二
東京都国分寺市東恋ケ窪１丁目280番地
株式会社日立製作所中央研究所内

(74)代理人　弁理士　小川　勝男

最終頁に続く

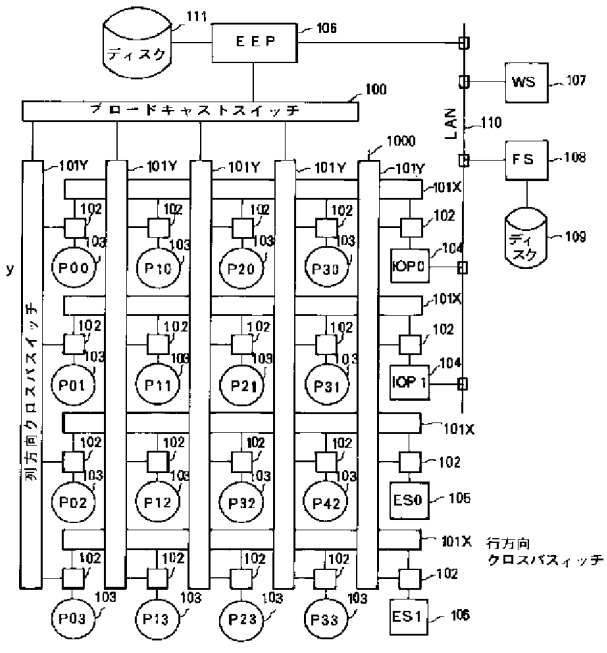
(54)【発明の名称】　並列プロセッサ

(57)【要約】　（修正有）

【目的】並列処理を実行するプロセッサの台数を減らすことなく、並列処理の管理機能や入出力機能を実現する並列プロセッサと、その実装方法を提供する。

【構成】２のべき乗＋１個のポートを持つクロスバスイッチ１０１を設け、そのクロスバスイッチに２のべき乗個となるようにプロセッサ１０３群を配置する。また各クロスバスイッチ１０１の残りに並列処理の管理機能や入出力機能を実行するプロセッサ１０４、１０５、１０６を配置し、又各プロセッサとクロスバスイッチ１０１の接続には乗り換えスイッチ１０２を設ける。これにより並列処理の実行は２のべき乗個のプロセッサで実行し、並列処理の管理機能や入出力機能はそれ以外のプロセッサでその処理とは独立に実行できるようにした。並列プロセッサの実装では、ある１つの次元のクロスバスイッチ、そのクロスバスイッチに接続するプロセッサ群、および、そのプロセッサ群の中の１個のプロセッサに接続する。

図１



【特許請求の範囲】
【請求項１】それぞれ並列処理プログラムを実行する複数の実行プロセッサと、それらを補助するための複数の補助プロセッサと、それらの実行プロセッサとそれらの補助プロセッサを接続するネットワークからなる並列プロセッサにおいて、
そのネットワークは、
それぞれ、並列にデータを転送可能な複数の部分ネットワークからなる複数群の部分ネットワークと、
互いに異なる部分ネットワークに属する複数の部分ネットワークにそれぞれ含まれる複数の入出力ポートをそれぞれ接続する複数の乗り換えスイッチとを有し、
該複数群の一つに属する部分ネットワークは、それぞれ２のべき乗プラス１個の入出力ポートを有し、
他の群の部分ネットワークの各々は、その群により定められた２のべき乗箇の入出力ポートを有し、
該複数の実行プロセッサは、該一つの群に属する複数の部分ネットワークの各々に含まれる２のべき乗個の入出力ポートの一つにそれぞれ接続された複数の乗り換えスイッチに接続され、
該複数の補助プロセッサは、該一つの群に属する複数の部分ネットワークの各々に含まれる該２のべき乗個の入出力ポート以外の一つの入出力ポートにそれぞれ接続された複数の乗り換えスイッチに接続されている並列プロセッサ。
【請求項２】各部分ネットワークは、クロスバスイッチからなる請求項１記載の並列プロセッサ。
【請求項３】該複数の補助プロセッサは、複数の入出力プロセッサを含む請求項１記載の並列プロセッサ。
【請求項４】該複数の補助プロセッサは、拡張記憶をそれぞれ有する複数の拡張記憶プロセッサを含む請求項１記載の並列プロセッサ。
【請求項５】該複数の補助プロセッサは、複数の入出力プロセッサと複数のフロントエンドプロセッサを含む請求項１記載の並列プロセッサ。
【請求項６】該複数の補助プロセッサは、少なくとも一つのフロントエンドプロセッサを含む請求項１記載の並列プロセッサ。
【請求項７】少なくとも１つのフロントエンドプロセッサと、
該他の群の内の一つの群に属する複数の部分ネットワークの各々に対応する入出力ポートおよび該フロントエンドプロセッサに接続された入出力ポートを有する他の一つの部分ネットワークとを更に有し、
該他の一つの群に属する複数の部分ネットワークの各々は、該他の部分ネットワークの一つの入力ポートに接続された一つの入出力ポートを更に有する請求項１記載の並列プロセッサ。
【請求項８】それぞれ並列処理プログラムを実行する複数の実行プロセッサと、それらを補助するための複数の補助プロセッサと、それらの実行プロセッサとそれらの補助プロセッサを接続するネットワークからなる並列プロセッサにおいて、
そのネットワークは、
それぞれ、２のｉ乗（ｉは、正の整数）箇の第１の座標軸の座標点群の１つに対応して設けられ、それぞれ２のｊ乗（ｊは、正の整数）プラス１箇の入出力ポートを有する、２のｉ乗箇の第１種のクロスバスイッチと、
それぞれ、２のｊ乗箇プラス１箇の第２の座標軸の座標点群の１つに対応して設けられ、それぞれ少なくとも２のｉ乗箇の入出力ポートを有する、２のｊ乗箇プラス１箇の第２種のクロスバスイッチと、
それぞれ該第１の座標軸の座標点群と該第２の座標軸の座標点群との異なる組の１つに対応して配置され、それぞれ対応する組の第１の座標軸の座標点に対応する１つの第１種のクロスバスイッチ内の１つの入出力ポートと、該対応する組の第２の座標軸の座標点に対応する１つの第２種のクロスバスイッチ内の１つの入出力ポートとに接続され、それらの間でデータの転送をする複数の乗り換えスイッチを有し、
該複数のプロセッサは、該複数の乗り換えスイッチのうち、該第１の座標軸の座標点群と該第２の座標点群の内の連続する２のｊ乗箇の座標点との異なる組の１つに対応してそれぞれ配列された第２群の乗り換えスイッチの対応する１つにそれぞれ接続されている並列プロセッサ。
【請求項９】並列処理プログラムを並列に実行する、複数の実行プロセッサと、それらを補助するための複数の補助プロセッサと、それらの実行プロセッサとそれらの補助プロセッサを接続するネットワークからなる並列プロセッサにおいて、
そのネットワークは、
それぞれ、２のｉ乗（ｉは、正の整数）箇の第１の座標軸の座標点群と２のｋ乗（ｋは、正の整数）箇の第３の座標軸の座標点群との異なる組の１つにそれぞれ対応して設けられ、それぞれ２のｊ乗（ｊは、正の整数）プラス１箇の入出力ポートを有する、２の（ｉプラスｋ）乗箇の第１種のクロスバスイッチと、
それぞれ、該第２の座標軸の座標点群と該第３の座標軸の座標点群との異なる組のそれぞれに対応して設けられ、それぞれ少なくとも２のｉ乗箇の入出力ポートを有する、２のｊ乗箇プラス１掛ける２のｋ乗箇の第２種のクロスバスイッチと、
それぞれ、該第１の座標軸の座標点群と該第２の座標軸の座標点群との異なる組のそれぞれに対応して設けられ、それぞれ少なくとも２のｋ乗箇の入出力ポートを有する、２のｉ乗掛ける（２のｊ乗プラス１）箇の第３種のクロスバスイッチと、
それぞれ該第１の座標軸の座標点群と該第２の座標軸の座標点群と該第３の座標軸の座標点群との異なる組の１

つに対応して配置され、それぞれ対応する組の第１の座標軸の座標点に対応する１つの第１種のクロスバスイッチ内の１つの入出力ポートと、該対応する組の第２の座標軸の座標点に対応する１つの第２種のクロスバスイッチ内の１つの入出力ポートと、該対応する組の第３の座標軸の座標点に対応する１つの第３種のクロスバスイッチ内の１つの入出力ポートとに接続され、それらの間でデータの転送をする複数の乗り換えスイッチを有し、
該複数のプロセッサは、該複数の乗り換えスイッチのうち、該第１の座標軸の座標点群と該第２の座標点群の内の連続する２のｊ乗箇の座標点と該第３の座標軸の座標点群との異なる組の１つに対応してそれぞれ配列された第１群の乗り換えスイッチの対応する１つにそれぞれ接続され、
該複数の補助プロセッサの少なくとも一部の複数の補助プロセッサは、該複数の乗り換えスイッチのうち、該第１の座標軸の座標点群と該連続する複数の第２の座標軸の座標点以外の第２の座標点と該第３の座標軸の座標点群との異なる組の１つにそれぞれ対応する第２群の乗り換えスイッチの対応する１つに接続されている並列プロセッサ。
【請求項１０】各第１種のクロスバスイッチと各第２種のクロスバスイッチは、互いに同じ回路からなる、請求項８または９記載の並列プロセッサ。
【請求項１１】それぞれｐ箇の第１の座標軸の座標点群の１つに対応して設けられ、それぞれｑ箇の入出力ポートを有する、ｐ箇の第１種のクロスバスイッチと、
それぞれｑ箇の第２の座標軸の座標点群の１つに対応して設けられ、それぞれｐ箇の入出力ポートを有する、ｑ箇の第２種のクロスバスイッチと、
それぞれ第１から第３の入出力ポートを有し、それらの間でデータの転送をするｐ箇掛けるｑ箇の乗換えスイッチからなり、
それぞれの乗換えスイッチは、該第１の座標軸の座標点群と該第２の座標軸の座標点群との異なる組の１つに対応して配列され、それぞれの第１の入出力ポートが対応する組の第１の座標軸の座標点群に対応する１つの第１種のクロスバスイッチ内の１つの入出力ポートに接続され、それぞれの第２の入出力ポートが対応する組の第２の座標軸の座標点群に対応する１つの第２種のクロスバスイッチ内の１つの入出力ポートに接続され、それぞれの第３の入出力ポートはデータ処理装置が接続され、
該ｐ箇の第１種のクロスバスイッチと該ｑ箇の第２種のクロスバスイッチとｐ箇掛けるｑ箇の乗換えスイッチは、相互に接続された実質的に同一構造の複数の実装単位上に構成され、
各実装単位は、
１つの第１種のクロスバスイッチと、
１つの第２種のクロスバスイッチと、
それぞれ該１つの第１種のクロスバスイッチのｐ箇の入出力ポートの対応する１つにその第１の入出力ポートに接続された、ｐ箇の乗換えスイッチと、
そのｐ箇の乗換えスイッチの内の１つの乗換えスイッチの第２の入出力ポートが上記１つの第２種のクロスバスイッチの１つの入出力ポートに接続され、
該実装単位は、そのｐ箇の乗換えスイッチの内の残りのｐ箇の乗換えスイッチの第２の入出力ポートがそれぞれ他の対応する１つの実装単位上の１つの第２種のクロスバスイッチの１つの入出力ポートに接続されているクロスバネットワーク。
【請求項１２】該ｐと該ｑはともに２のｉ乗プラス１（ｉは正の整数）である請求項１１記載のクロスバネットワーク。
【請求項１３】各実装単位上に該複数の装置の１部の装置がそれぞれ、その上の該ｐ箇の乗換えスイッチのいずれか１つに含まれている第３の入出力ポートに接続されて、実装されている請求項１１記載の並列プロセッサ。
【請求項１４】それぞれ並列処理プログラムを実行する複数の実行プロセッサと、それらの実行プロセッサを接続するネットワークからなる並列プロセッサにおいて、
少なくとも１つのフロントエンドプロセッサと、
そのネットワークは、
それぞれ、並列にデータを転送可能な複数の部分ネットワークからなる複数群の部分ネットワークと、
互いに異なる部分ネットワークに属する複数の部分ネットワークにそれぞれ含まれる複数の入出力ポートをそれぞれ接続する複数の乗り換えスイッチとを有し、
該複数の実行プロセッサは、該一つの群に属する複数の部分ネットワークの各々に含まれる２のべき乗個の入出力ポートの一つにそれぞれ接続された複数の乗り換えスイッチに接続され、
該他の群の内の一つの群に属する複数の部分ネットワークの各々に対応する入出力ポートおよび該フロントエンドプロセッサに接続された入出力ポートを有する他の一つの部分ネットワークとを更に有し、
該他の一つの群に属する複数の部分ネットワークの各々は、該他の部分ネットワークの一つの入力ポートに接続された一つの入出力ポートを更に有する並列プロセッサ。
【請求項１５】各部分ネットワークは、クロスバスイッチからなる請求項１４に記載の並列プロセッサ。
【発明の詳細な説明】
【０００１】
【産業上の利用分野】本発明は並列プロセッサにおけるプロセッサの結合方式に関わり、特に並列処理の高速化に好適な並列プロセッサの構成と、その実装方法に関する。
【０００２】
【従来の技術】従来、並列処理プログラムを実行する複数のプロセッサ（以下、実行プロセッサと呼ぶ）と、プ

ロセッサ間のデータ転送を可能にするネットワークにより構成される並列プロセッサは、たとえば特開昭６３－１２４１６２号（以下、第１の従来例と呼ぶ）に示されている。この並列プロセッサにおいては縦２のｍ乗行と横２のｎ乗列（ｍ，ｎは正の整数）からなる２次元アレイ状に実行プロセッサを配置し、このアレイの各行（Ｘ方向）、各列（Ｙ方向）ごとにクロスバスイッチを設ける。各実行プロセッサはＸ／Ｙ方向のそれぞれの１つのクロスバスイッチに結合される。このため各実行プロセッサには１つのＸ方向のクロスバスイッチと１つのＹ方向のクロスバスイッチを結合するために２つの入出力ポート（以下、単にポートと呼ぶ）を有している。以下、このような各次元ごとに設けたクロスバスイッチネットワークを部分クロスバネットワークと呼ぶ。

【０００３】一方ＮＣＵＢＥ社のＮＣＵＢＥ（以下、第２の従来例と呼ぶ）ではハイパキューブ構成のネットワークに入出力プロセッサを結合している（例えば、「３２ビットマイクロプロセッサの全容　企業戦略・技術・市場動向」２３６、２３８ページ、日経マグローヒル社発行、１９８６年１２月）。これを実現するために各ノードプロセッサでは、ハイパキューブの次元数分に１つ加えた数のポートを持っている。ハイパキューブの次元数のポートには各ノードプロセッサを結合し、残りの１つには入出力プロセッサを結合する。即ち各ノードプロセッサは、補助プロセッサと接続するための専用のポートを有している。

【０００４】また従来、部分クロスバネットワークを構成する並列プロセッサの実装については、例えばザ・セカンド・シンポジューム・オン・ザ・フロンティア・オブ・マッシブリィ・パラレル・コンピューテーション（１９８８年１０月）第４６３ページから第４６６ページ（The 2nd Syposium on the Frontiers of Massively Parallel Computation,October 10-12 ,1988,pp463-466）、および電子情報通信学会技術研究報告、ＣＰＳＹ８９－４５～５８、１９８９年３月、第３９ページから第４４ページ（それぞれ第３、第４の従来例と呼ぶ）に示されている。

【０００５】第３の従来例では２次元の部分クロスバネットワークで構成された並列プロセッサを、ローカルクロスバネットワークと呼ばれる１つのクロスバスイッチとそれに接続されるプロセッサ群を１つのウェハに実装する。またグローバルクロスバネットワークと呼ばれるもう１つのクロスバスイッチは前記ウェハとは別に、専用のＬＳＩに実装する。このウェハと専用のＬＳＩを複数個、組み合せて並列プロセッサを構成する。

【０００６】第４の従来例では３次元の部分クロスバネットワークで構成された並列プロセッサを、１つの面に対応するＸ方向とＹ方向のクロスバスイッチ群とそれに接続されるプロセッサ群を１つの筐体に実装する。またＺ方向のクロスバスイッチ群は専用の筐体に実装する。これらの筐体を複数個、組み合せて並列プロセッサを構成する。

【０００７】

【発明が解決しようとする課題】各実行プロセッサをＸ方向とＹ方向のクロスバスイッチに結合した上記第１の従来技術ではプロセッサの結合方式についてのみ開示されている。しかし並列処理を実行するためにはプロセッサ群への処理分配などの管理機能や入出力機能も不可欠であるが、これを行なう補助プロセッサについては言及していない。また、この補助プロセッサを上記プロセッサ群へ接続するためのネットワークを示していない。上記第２の従来例技術のように各実行プロセッサと補助プロセッサを直結したのでは、各実行プロセッサに新規に１つのポートを設け必要があり、かつ、そのポートでのメッセージ転送制御回路が複雑とならざるを得ない。また、この従来技術では補助プロセッサと各実行プロセッサを接続するネットワークを別に設けているが、この別のネットワークは実行プロセッサの台数の増大とともに複雑化せざるを得ない。

【０００８】また第３、４の従来技術による並列プロセッサの実装においては、プロセッサの実装とは別に、クロスバスイッチ群を実装するための専用のＬＳＩ、筐体などが必要である。これによればプロセッサ台数が多くなると、プロセッサとクロスバスイッチを接続するためのインタフェースが一ヶ所に集中し、そのインタフェースが多くなる。このためクロスバスイッチ群を実装する専用のＬＳＩ、筐体などの物理ピン数を越えることになり、そのインタフェースが接続できない場合がある。

【０００９】本発明の第１の目的は補助プロセッサを実行プロセッサに接続する比較的構造の簡単なネットワークを有する並列プロセッサを提供することにある。

【００１０】本発明の第２の目的はプロセッサとネットワークを接続するためのインタフェースを一ヶ所に集中させない並列プロセッサの実装方法を提供することにある。

【００１１】

【課題を解決するための手段】上記第１の目的を達成するために、本発明では並列処理の管理機能や入出力機能を実行するための補助プロセッサを、任意の実行プロセッサ間でデータ転送を行なう部分クロスバネットワークに接続する。各次元のクロスバスイッチとして２のべき乗個のポートに１個のポートを追加したポート数（２のべき乗＋１個）のポートを持つ部分クロスバスイッチを用いる。各次元が２のべき乗個となるようにｎ次元直方体に実行プロセッサ群を配置し、それらのクロスバスイッチに乗換えスイッチを介して、また各クロスバスイッチの残りの１ポートのいずれかに並列処理の管理機能や入出力機能を実行する補助プロセッサを乗換えスイッチを介して接続する。

【００１２】上記第２の目的を達成するために、本発明

では、ある１つの次元のクロスバスイッチ、そのクロスバスイッチに接続するプロセッサ群、および、そのプロセッサ群の中の１個のプロセッサに接続する、かつ、上記とは別の次元の全てのクロスバスイッチを、１つの実装単位とする。

【００１３】

【作用】並列処理は、従来通り、各次元が２のべき乗個となるようにｎ次元直方体に配置した実行プロセッサ群で実行できる。並列処理を実行する実行プロセッサからは、プロセッサ間データ転送と同様にして補助プロセッサに対して通信が可能である。これにより各実行プロセッサにポートを追加することなく、また補助プロセッサを結合するためのネットワークの追加はごくわずかである。このため並列プロセッサの実装規模を小さくできる。

【００１４】並列プロセッサの実装においては、クロスバスイッチ群を実装するための専用のＬＳＩ、筐体などは必要なく、この実装単位だけの組合せで可能である。これによりプロセッサとネットワークを接続するインタフェースを一ヶ所に集中させることなく実装できる。

【００１５】

【実施例】以下、本発明の一実施例を図を用いて説明する。図１に並列プロセッサのシステム構成を示す。本並列プロセッサにおいては行方向または列方向のクロスバスイッチ１０１（行方向クロスバスイッチを１０１ｘ、列方向クロスバスイッチを１０１ｙと呼び、これらを区別しない場合には単に１０１と呼ぶ）、行方向クロスバスイッチ１０１ｘと列方向クロスバスイッチ１０１ｙ間の乗り換えを行なう複数の乗り換えスイッチ１０２、ｉ行ｉ列に配列された並列処理を実行するプロセッサ（以下、計算クラスタと呼ぶ）１０３、入出力機能を実行する入出力プロセッサ（以下、入出力クラスタと呼ぶ）１０４、各計算クラスタ１０３で共有するデータや中間結果などを保持する拡張記憶プロセッサ（以下、ＥＳクラスタと呼ぶ）１０５、並列処理の管理機能を実行するフロントエンドプロセッサ（以下、ＦＥＰと呼ぶ）１０６、一般的なユーザ環境を持つワークステーション（以下、ＷＳと呼ぶ）１０７、ユーザファイルを管理するファイルサーバ（以下、ＦＳと呼ぶ）１０８、ユーザファイルを保持するディスク装置１０９、ＷＳ１０７などのコンピュータ間の通信を高速に行なうローカルエリアネットワーク（以下、ＬＡＮと呼ぶ）１１０により構成する。本実施例では、各クロスバスイッチ１０１は、２のｉ乗プラス１個のポートを有し、同じ構造を有し、さらに計算クラスタ１０３は２のｉ乗行２のｉ乗列に配列されるのが特徴である。

【００１６】図１では一例として計算クラスタ１０３の台数を合計１６台とし、それを４×４の２次元アレイ状に配置したものを示す。従って各クロスバスイッチは５個のポートを有する。しかし本発明は任意の値のｉに対して有効である。

【００１７】各計算クラスタ１０３は、行方向（以下、Ｘ方向と呼ぶ）と列方向（Ｙ方向と呼ぶ）の各クロスバスイッチ１０１に乗換えスイッチ１０２を介して接続する。図１では入出力クラスタ１０４とＥＳクラスタ１０５はそれぞれ２台づつあり、Ｘ方向のクロスバスイッチ１０１ｘの残りの１つのポートに接続する。ＦＥＰ１０６は計算クラスタ１０３を接続するＹ方向の各クロスバスイッチ１０１ｙの残りの１つのポートにブロードキャストスイッチ１００により接続される。ただし入出力クラスタ１０４とＥＳクラスタ１０５とが接続されたＹ方向のクロスバスイッチ１０１ｙ（図の右端のもの）の残りの１つのポート１０００は、このクロスバスイッチ１００には接続されない。これはブロードキャストスイッチ１００のポートの数をクロスバスイッチ１０１と同じにしたため、このポート１０００を接続するためのポートが存在しないからである。ここでは、このブロードキャストスイッチ１００はクロスバスイッチ１０１と同じ構成である。各クロスバスイッチ１０１と各入出力クラスタ１０４、ＥＳクラスタ１０５の接続にもそれらの間で乗り換えを行なう乗り換えスイッチ１０２を設ける。結局、本実施例では、２次元空間のＸ座標が０、１、２、３、Ｙ座標が０、１、２、３からなる点群に対応して計算クラスタ１０３が配置される。各Ｙ座標０～３の各点に対応して１つのＸ方向のクロスバスイッチが設けられ、各Ｘ座標に対応して１つのＹ方向のクロスバスイッチが設けらている。各入出力クラスタ１０４とＥＳクラスタ１０５がＸ＝４、Ｙ＝０～３の座標点群に対応して設けられていることになる。以下、これらのクロスバスイッチおよび乗換えスイッチにより構成されるネットワークをクロスバネットワークと呼ぶ。

【００１８】本並列プロセッサは、ＷＳ１０７などで作成された計算依頼をＬＡＮ１１０を介してＦＥＰ１０６で受け付ける。ＦＥＰ１０６にＷＳ１０７の機能も持ち、ＦＥＰ１０６で直接、計算依頼を作成することも可能である。ＦＥＰ１０６では、その計算依頼をスプーリングし、その計算依頼で指示されたプログラムの実行が可能なクラスタにそれを割当るなどの処理を行なう。また、これらのためにＦＥＰ１０６ではディスク装置１１１を保持する。各計算クラスタ１０３で実行するプログラムやデータにおいては、まずＦＥＰ１０６が各計算クラスタ１０３に対してプログラムをロードするための情報を配った後、各計算クラスタ１０３にプログラムの実行開始を指示する。そして各計算クラスタ１０３がそれを受け取ると、計算クラスタ１０３内のオペレーティングシステムにより、クロスバネットワークを介して、いずれかの入出力プロセッサ１０４に依頼し、ＦＳ１０８からその入出力プロセッサを介してプログラムやデータをロードする。ロードが終了すると、各計算クラスタ１０３はプログラムの実行を開始する。

【００１９】各計算クラスタ１０３間でデータ転送が必要な時はクロスバネットワークを介して行なう。たとえば、計算クラスタＰ００からＰ３３へのデータ転送では、まずＰ００から送出されたメッセージはＰ００に接続される乗り換えスイッチ１０２からＸ方向のクロスバスイッチ１０１ｘ（図１の最上段の１０１ｘ）へ送られる。そのＸ方向のクロスバスイッチ１０１ｘを通ったメッセージは、Ｐ３０に接続される乗り換えスイッチ１０２によりＹ方向のクロスバスイッチ１０１ｙ（図１の最右側から２番目の１０１ｙ）に乗り換えられる。そして、Ｙ方向のクロスバスイッチ１０１ｙからＰ３３に接続される乗り換えスイッチ１０２に送られ、Ｐ３３に届く。また同様に計算クラスタ１０３が計算の途中で入出力を必要とする時は入出力クラスタ１０４にクロスバネットワークを介して依頼する。依頼された入出力クラスタ１０４では依頼された内容に従い動作し、応答が必要な場合は、その結果を依頼元の計算クラスタ１０３に返す。またＥＳクラスタ１０５では、各計算クラスタ１０３が起動される前にＦＥＰ１０６によりデータがロードされ、各計算クラスタ１０３は実行時にクロスバネットワークを介してそのデータをアクセスする。また計算クラスタ１０３が実行時に計算結果などを直接書き込んでもよい。

【００２０】図２に計算クラスタの構成を示す。計算クラスタ１０３は、一般的なプロセッサにクロスバネットワークとのインタフェースを持つように構成されていればよく、たとえば複数の計算処理を実行するプロセッシングユニット２０１、各計算クラスタ１０３間とのデータ転送を実行するネットワークインタフェースアダプタ２０２、プログラムやデータを記憶するメモリ２０３、計算クラスタ内のオペレーティングシステム（以下、ＯＳ）のブートデバイスやページングおよびスワップデバイスなどとして用いるディスク装置２０５および、そのディスク装置２０５との入出力インタフェースを持つ入出力アダプタ２０４により構成される。ここでは計算クラスタ１０３がマルチプロセッサ構成をとるようにプロセッシングユニット２０１が複数台あるが、１台であってもよい。

【００２１】プロセッシングユニット２０１の処理において、計算クラスタ間のデータ転送要求、入出力要求、拡張記憶へのアクセス、ＦＥＰへの通信などがある場合は、ネットワークインタフェースアダプタ２０２へその処理を依頼する。ネットワークインタフェースアダプタ２０２ではこれらの要求を受け付けると、それに必要なメッセージを組立て、クロスバネットワークへ送り出す。またネットワークインタフェースアダプタ２０２ではクロスバネットワークから送られてくるメッセージを受け付け、それに対応した処理を行なう。その処理はたとえばプロセッシングユニット２０１にメッセージを渡すために割込みをかける、メッセージで指定された転送データをメモリ２０３に書き込むなどである。

【００２２】図３に入出力クラスタの構成を示す。入出力クラスタ１０４は、計算クラスタ１０３と同じようにネットワークインタフェースアダプタ２０２、メモリ２０３、入出力アダプタ２０４、ディスク装置２０５を持ち、さらに入出力を制御する入出力制御プロセッサ３０１、ＬＡＮ１１０のインタフェースを持つＬＡＮアダプタ３０２により構成する。

【００２３】入出力クラスタ１０４では、計算クラスタ１０３からの入出力要求メッセージをネットワークインタフェースアダプタ２０２で受け付けると、そのメッセージを入出力制御プロセッサ３０１に渡す。入出力制御プロセッサ３０１ではそのメッセージに従い、入出力アダプタ２０４、ＬＡＮアダプタ３０２を介してディスク２０５に対して入出力動作を実行する。入出力動作後、計算クラスタ１０３への応答が必要な場合はネットワークインタフェースアダプタ２０２にその処理を依頼する。また、この処理以外にも入出力クラスタ１０４ではファイルサーバの機能も有する。

【００２４】図４にＥＳクラスタの構成を示す。ＥＳクラスタ１０５はＥＳ制御プロセッサ４０１、ネットワークインタフェースアダプタ２０２、メモリ２０３、および拡張記憶（ＥＳ）４０２により構成される。

【００２５】ＥＳ制御プロセッサ４０１は、後述するロック要求メッセージ、ＥＳ読出し要求メッセージなどの処理を行なう。また計算クラスタ１０３への応答がある場合は、ネットワークインタフェースアダプタ２０２へその処理を依頼する。ネットワークインタフェースアダプタ２０２では要求を受け付けると、それに必要なメッセージを組立て、クロスバネットワークへ送り出す。またネットワークインタフェースアダプタ２０２ではクロスバネットワークから送られてくるメッセージを受け付け、それに対応した処理を行なう。その処理はたとえばＥＳ制御プロセッサ４０１にメッセージを渡すために割込みをかける、メッセージで指定されたＥＳ書き込みデータをＥＳ４０２に書き込むなどである。またメモリ２０３はＥＳ制御プロセッサ４０１で実行するプログラムやデータなどを格納するものである。このＥＳ制御プロセッサ４０１は、一般的なマイクロプロセッサとメモリ制御ユニットで構成されてもよい。

【００２６】次にメッセージの転送について説明する。図５に計算クラスタ間データ転送のメッセージフォーマットを示す。このメッセージは、メッセージを受け取るクラスタを表わす受信クラスタ番号５０１、ネットワークの転送方法を表わす転送制御５０２、メッセージの種別、受信クラスタ内での処理方法を表わすデータ転送コマンド５０３、受信クラスタに転送データ５０７が届いたことを知らせるためのデータ識別子アドレス５０４、受信クラスタに転送データ５０７が届いた時にその転送データ５０７をメモリ２０３に書き込むための転送デー

タ受信アドレス５０５、転送データ５０７の長さを示す転送データ長５０６および転送データ５０７により構成される。

【００２７】各プロセッシングユニット２０１で実行されているプログラムがメッセージを転送する時には、プロセッシングユニット２０１はメッセージの各フィールドに設定する値をネットワークインタフェースアダプタ２０２に渡し、さらにメッセージ生成および送出の開始を指示する。ここで受信クラスタ番号５０１は図６に示すようにＸ方向／Ｙ方向のクロスバスイッチのポートを決定するＸ系クロスバ番号とＹ系クロスバ番号で構成される。また各計算クラスタ間のデータ転送における送受信クラスタ番号は、プログラムあるいはシステムからは一般に順序付けられた番号にした方が、指定しやすい。たとえば１６台の計算クラスタ構成の場合は０から１５の番号が付けらる。しかし実際のハードウェアで使用するクラスタ番号は異なる。これは１６台の計算クラスタ構成の場合、計算クラスタを指定するには４ビットであれば十分であるが、２のべき乗＋１個のポートを持つクロスバスイッチを用いたことにより、ハードウェアとしてはＸ系クロスバ番号／Ｙ系クロスバ番号にそれぞれ３ビットづつで合計６ビットが必要になる。これによると各計算クラスタは順序付けられた番号にならない。このため図１２に示すような計算クラスタの番号変換を行なう必要がある。この変換は、たとえば計算クラスタ内のネットワークインタフェースアダプタ２０２で実現する。またＦＥＰ１０６、入出力クラスタ１０４の番号変換についても同図に示すように同様に実現される。

【００２８】計算クラスタより送出されたメッセージは、クロスバスイッチ１０１および乗り換えスイッチ１０２がメッセージ内の受信クラスタ番号５０１および転送制御５０２を用いてルーティングを行ない、目的の受信クラスタに到着する。ここで転送制御５０２は図７に示すようにそのメッセージを全てのクラスタに転送するか否かを示すブロードキャスト制御とメッセージを転送する時にＸ方向あるいはＹ方向のどちらからを先に送出するかを決定する転送順序制御により構成される。この転送順序制御はクロスバネットワーク内のルーティングをたとえば、Ｘ方向からＹ方向に限る固定ルーティングとし、必ずその方向の転送だけになれば必要ないが、各クラスタからＦＥＰ１０６へのデータ転送のように構成次第で、Ｙ方向からＸ方向へメッセージを転送することが必要になるため転送順序制御を設けている。

【００２９】メッセージが届いた受信クラスタでは、ネットワークインタフェースアダプタ２０２がそのメッセージのデータ転送コマンド５０３をみて、転送データ５０７を転送データ受信アドレス５０５に従いメモリ２０３に書き込む。また受信した転送データ５０７をメモリ２０３に書き込んだことを知らせるために、データ識別子アドレス５０４に従いメモリ２０３にその情報を書き込む。受信クラスタ内のプロセッシングユニット２０１はデータ識別子アドレスによりメモリ２０３を監視し、転送データ５０７が届いたことを知る。

【００３０】次にブロードキャストのメッセージ転送について説明する。各乗換えスイッチ１０２では対応する計算クラスタ１０３から転送されてきたメッセージ内のブロードキャスト制御がブロードキャストを示していることを知ると、まず、そのスイッチに接続されたＸ方向クロスバスイッチ１０１ｘにそのメッセージを転送する。上記Ｘ方向クロスバスイッチ１０１ｘでは、それに接続される全ての乗換えスイッチ１０２に対して、転送可能であることを確認してからそのメッセージを転送する。Ｘ方向クロスバスイッチ１０１ｘからブロードキャストのメッセージが転送されてきた乗換えスイッチ１０２の各々では、次にそのスイッチに接続されたＹ方向クロスバスイッチ１０１ｙにそのメッセージを転送する。以下同様にＹ方向クロスバスイッチ１０１ｙから乗換えスイッチ１０２にブロードキャストのメッセージを転送する。この乗換えスイッチ１０２では、このメッセージをそれに接続される計算クラスタ１０３に転送する。このようにしてブロードキャストのメッセージ転送を実行する。

【００３１】次にＦＥＰ１０６と計算クラスタ１０３とのメッセージ転送について説明する。図１では、ＦＥＰ１０６と各計算クラスタ１０３とのメッセージ転送を効率よく行うためにブロードキャストスイッチ１００を１つ新たに設けている。これによりブロードキャストスイッチ１００と一群のＹ方向のクロスバスイッチ１０１ｙを経由するだけで、ＦＥＰ１０６といずれの計算クラスタ１０３との間でメッセージ転送が可能である。

【００３２】たとえば図１のシステム構成における計算クラスタ１０３のＰ００からＦＥＰ１０６へのメッセージ転送では、計算クラスタ１０３間のメッセージ転送と同様にしてネットワークインタフェースアダプタ２０２からメッセージが送出される。このときのメッセージ内の受信クラスタ番号は、図１２のクラスタ番号変換表で示すようにＦＥＰ１０６を示すＸ系クロスバ番号＝４、Ｙ系クロスバ番号＝４に変換されている。またメッセージ内の転送順序制御は、図１０に示すようにＹ方向クロスバスイッチからＸ方向クロスバスイッチへメッセージを転送するように、１に設定されている。計算クラスタ内のネットワークインタフェースアダプタ２０２から送出されたメッセージは、まずＰ００に接続される乗換えスイッチ１０２に送出される。その乗換えスイッチ１０２では、メッセージ内の転送順序制御により、それに接続されるＹ方向のクロスバスイッチ１０１ｙに転送する。そのＹ方向のクロスバスイッチ１０１ｙでは、図８で示すルーティング情報（詳細については後述する）に従い、メッセージ内の受信クラスタ番号のＹ系クロスバ番号に一致するポート番号８０４のポートにそのメッセ

ージを送出する。この場合、このポートにはブロードキャストスイッチ１００が接続されている。ブロードキャストスイッチ１００では、メッセージ内の受信クラスタ番号のＸ系クロスバ番号に一致するポート番号８０４のポートにそのメッセージを送出することにより、ＦＥＰ１０６へのメッセージ転送が行なわれる。

【００３３】ＦＥＰ１０６から計算クラスタ１０３のＰ００へのへのメッセージ転送では、計算クラスタ１０３間のメッセージ転送と同様にしてネットワークインタフェースアダプタ２０２からメッセージが送出される。このときのメッセージ内の受信クラスタ番号は、図１２のクラスタ番号変換表で示すように計算クラスタ１０３のＰ００を示すＸ系クロスバ番号＝０、Ｙ系クロスバ番号＝０に変換されている。またメッセージ内の転送順序制御は、図１０に示すようにＸ方向クロスバスイッチからＹ方向クロスバスイッチへメッセージを転送するように、０に設定されている。ＦＥＰ１０６のネットワークインタフェースアダプタ２０２から送出されたメッセージは、まずブロードキャストスイッチ１００に送出される。そのブロードキャストスイッチ１００では、図８で示すルーティング情報（詳細については後述する）に従い、メッセージ内の受信クラスタ番号のＸ系クロスバ番号に一致するポート番号８０４のポートにそのメッセージを送出する。この場合、このポートにはＹ方向のクロスバスイッチ１０１ｙが接続されている。そのＹ方向のクロスバスイッチ１０１ｙでは、メッセージ内の受信クラスタ番号のＹ系クロスバ番号に一致するポート番号８０４のポートにそのメッセージを送出する。このポートには、Ｐ００に接続される乗換えスイッチ１０２が接続されており、これにより計算クラスタ１０３のＰ００へのメッセージ転送が行なわれる。

【００３４】ＦＥＰ１１０６から全計算クラスタ１０３へのブロードキャスト転送では、計算クラスタ１０３間のメッセージ転送と同様にしてネットワークインタフェースアダプタ２０２からメッセージが送出される。このときのメッセージ内の受信クラスタ番号は何であってもよい。またメッセージ内の転送順序制御は、図１０に示すようにＸ方向クロスバスイッチからＹ方向クロスバスイッチへメッセージを転送するように０に設定されている。さらにメッセージ内のブロードキャスト制御がブロードキャストを示している。ＦＥＰ１０６のネットワークインタフェースアダプタ２０２から送出されたメッセージは、まずブロードキャストスイッチ１００に送出される。そのブロードキャストスイッチ１００では、ブロードキャストが指定されているため、全ポートに送出可能であることを確認して、そのメッセージを送出する。各々のＹ方向のクロスバスイッチ１０１ｙでは、ブロードキャストのメッセージが転送されてくると、それに接続される全ての乗換えスイッチ１０２に対して、メッセージが転送可能であることを確認して各乗換えスイッチ１０２に転送する。

【００３５】このようにＦＥＰ１１０６から全計算クラスタ１０３へのブロードキャストでは、計算クラスタ間のそれとは異なり、Ｘ方向のクロスバスイッチ１０１ｘを介することなく、ブロードキャストスイッチ１００と一群のＹ方向のクロスバスイッチ１０１ｙを経由するだけで、全ての計算クラスタ１０３にメッセージをブロードキャスト転送が可能である。また各計算クラスタ１０３からＦＥＰ１０６にメッセージを転送する場合にも、ブロードキャストスイッチ１００を使うことによりＸ方向のクロスバスイッチ１０１ｘを介さずに転送できるので、ネットワーク内で実行されている他の計算クラスタ間のメッセージ転送の妨げを少なくできる。もしＦＥＰ１０６がたとえば図１の右上端の入出力プロセッサ１０４の位置にあった場合には、各計算クラスタ１０３から送出された複数のメッセージがそのＦＥＰ１０６が接続されるＸ方向のクロスバスイッチ１０１ｘに集中することになり、そのＸ方向のクロスバスイッチ１０１ｘを使用する他のメッセージ転送が妨げられるが、本実施例では、このような問題はない。

【００３６】次にこのネットワークを構成するハードウェアの実現方法について説明する。図８にメッセージの転送に必要なルーティング情報を示す。このルーティング情報はシステムが立ち上がった時に予め設定される。各クラスタ１０３、１０４、１０５およびＦＥＰ１０６には各クラスタの自クラスタ番号８０１を持つ。この番号８０１は送信クラスタが応答を必要とする時など、メッセージにそれを追加する。受信クラスタではそれに対応するメッセージを送り返す時、その番号が受信クラスタ番号として使用される。また各乗り換えスイッチ１０２にもそれに接続される各クラスタの自クラスタ番号８０２を持つ。この番号８０２はメッセージ内の受信クラスタ番号５０１と比較される。乗り換えスイッチ１０２は、その結果によりクラスタやクロスバスイッチにメッセージを渡す。各クロスバスイッチ１０１には各クロスバスイッチ１０１がＸ方向／Ｙ方向かの次元を示すクロスバスイッチ次元情報８０３および、各クロスバスイッチ１０１のそれぞれのポートがどこに接続されているかを示すポート番号８０４を持つ。クロスバスイッチ次元情報８０３はクロスバスイッチ１０１内のルーティングにおいて、メッセージ内の受信クラスタ番号５０１のＸ系クロスバ番号／Ｙ系クロスバ番号のどちらを切り出すかを決定する。またポート番号８０４は受信クラスタ番号５０１から切り出されたＸ／Ｙ系クロスバ番号と比較され、一致するポートへメッセージを送出するために使用される。

【００３７】図９に乗り換えスイッチの構成を示す。乗り換えスイッチ１０２は、乗り換えスイッチ１０２に接続されるクラスタ、クロスバスイッチから送出されるメッセージを一時保持するファーストーイン・ファースト

ーアウト・メモリ（以下、ＦＩＦＯ）９０１、ＦＩＦＯ９０１に保持されたメッセージを取り出し、そのメッセージの受信クラスタ番号５０１と転送制御５０２により転送先を決定するルーティング制御９０２および各ルーティング制御９０２からの要求に対して優先順位を決定し、各ポートにメッセージを送出するメッセージ送出制御９０３を各ポート対応に設け、それらと自クラスタ番号８０２により構成する。

【００３８】ＦＩＦＯ９０１は読み出しと書き込み動作が非同期に行なえる。これには市販されているＦＩＦＯ用ＬＳＩを用いてもよい。ルーティング制御９０２はＦＩＦＯ読み出し制御９２１と転送先決定回路９２２により構成する。ＦＩＦＯ読み出し制御９２１では、ＦＩＦＯ９０１から出力されるＥＭＰＴＹ信号９１１が０であることを確認して、リードストローブ９１２を送出し、ＦＩＦＯ９０１に保持されたメッセージを読みだす。このＥＭＰＴＹ信号９１１はＦＩＦＯ９０１に１個もデータが入ってない時は１で、それ以外は０である。

【００３９】転送先決定回路９２２では、図１０に示すように、読みだされたメッセージ内の受信クラスタ番号５０１のＸ系クロスバ番号、Ｙ系クロスバ番号がそれぞれ各乗り換えスイッチ１０２に保持されている自クラスタ番号８０２のＸ系クロスバ番号、Ｙ系クロスバ番号と比較される。この比較結果に基づきメッセージ内の転送制御５０２の転送順序制御の内容によりＸ方向から先に送出するか，あるいはＹ方向から先に送出するかを決定する。この転送先決定回路９２２の動作は、たとえば転送順序制御がＸ方向からＹ方向への転送が指定されている、即ち転送順序制御が０であると、Ｘ系クロスバ番号の比較結果が０、即ち不一致で、かつＹ系クロスバ番号の比較結果が０の場合Ｘ系クロスバスイッチに転送、Ｘ系クロスバ番号の比較結果が０で、かつＹ系クロスバ番号の比較結果が１、即ち一致の場合Ｘ系クロスバスイッチに転送、Ｘ系クロスバ番号の比較結果が１で、かつＹ系クロスバ番号の比較結果が０の場合Ｙ系クロスバスイッチに転送、Ｘ系クロスバ番号の比較結果が１で、かつＹ系クロスバ番号の比較結果が１の場合クラスタに転送することを示す。そして転送先決定回路９２２では決定した転送先に、それに対応する信号９２３によりメッセージ送出制御９０３に送出することを指示する。

【００４０】メッセージ送出制御９０３はプライオリティ制御９３１とセレクタ９３２により構成される。プライオリティ制御９３１では、各ルーティング制御からの信号９２３を受け取り、それらの間で送出する優先順位を決め、それに対して送出を行なう。送出の制御は、メッセージ送出制御９０３が送出先に乗り換えスイッチと同様に設けられているＦＩＦＯから出力されるＦＵＬＬ信号９０４が０であることを確認して、ライトストローブ９０５を送出し、送出先のＦＩＦＯにメッセージを書き込む。このＦＵＬＬ信号９０４は送出先ＦＩＦＯが一杯で１個のデータも受け取れない時は１で、それ以外は０である。またプライオリティ制御９３１では一回の送出でメッセージの転送が終わらない場合、送り出すごとに信号９３３により対応するＦＩＦＯ読み出し制御９２１に対してＦＩＦＯ９０１から次のデータを読みだすことを指示する。

【００４１】次にこの乗換えスイッチ１０２におけるブロードキャストのメッセージ転送の動作について説明する。ここではブロードキャストは、前述したように送出元計算クラスタ１０３、乗換えスイッチ１０２、Ｘ方向クロスバスイッチ１０１ｘ、乗換えスイッチ１０２、Ｙ方向クロスバスイッチ１０１ｙ、乗換えスイッチ１０２、全計算クラスタ１０３の順番でメッセージの転送が行なわれる。

【００４２】まず計算クラスタ１０３から送出されたメッセージは、乗換えスイッチ１０２に送出される。そして乗換えスイッチ１０２のルーティング制御９０２が、メッセージ内のブロードキャスト制御によりブロードキャストが指定されているかを判定する。ブロードキャストが指定されていると、ルーティング制御９０２は出力ポート番号レジスタ９２４に保持されているポート番号に従い、そのブロードキャストのメッセージをＸ方向のクロスバスイッチ１０１ｘが接続されるポートに出力する。

【００４３】ここで出力ポート番号レジスタ９２４の内容は、ブロードキャストのメッセージをどこのポートに出力するかを示す情報である。この場合の出力ポート番号レジスタ９２４の内容は、計算クラスタからブロードキャストのメッセージを受け取るルーティング制御９０２はＸ方向クロスバスイッチ１０１ｘに出力するように、Ｘ方向クロスバスイッチ１０１ｘからブロードキャストのメッセージを受け取るルーティング制御９０２はＹ方向クロスバスイッチ１０１ｙに出力するように、Ｙ方向クロスバスイッチ１０１ｙからブロードキャストのメッセージを受け取るルーティング制御９０２は計算クラスタに出力するように設定されている。また同様にＸ方向のクロスバスイッチ１０１ｘから送出されたブロードキャストのメッセージは、乗換えスイッチ１０２のルーティング制御９０２が判定を行ない、ブロードキャストが指定されているとルーティング制御９０２は出力ポート番号レジスタ９２４に保持されているポート番号に従い、Ｙ方向のクロスバスイッチ１０１ｙが接続されるポートに出力する。また同様にＹ方向のクロスバスイッチ１０１ｙから送出されたブロードキャストのメッセージは、乗換えスイッチ１０２のルーティング制御９０２が判定を行ない、ブロードキャストが指定されているとルーティング制御９０２は出力ポート番号レジスタ９２４に保持されているポート番号に従い、計算クラスタが接続されるポートに出力する。ブロードキャストのメッセージ転送では、この動作以外は他の転送制御と同じで

ある。図１１にクロスバスイッチの構成を示す。クロスバスイッチ１０１は、それに接続される乗り換えスイッチ１０２などから送出されるメッセージを一時保持し、それを要求に応じて取り出しを行なうメッセージ受信制御１１０１および、そのメッセージの受信クラスタ番号５０１と転送制御５０２により自ポートへ出力するかを決定し、各メッセージ受信制御１１０１からの要求の優先順位を決定し、各ポートにメッセージを送出するメッセージ送出制御１１０２を各ポート対応に設け、構成する。

【００４４】メッセージ受信制御１１０１はメッセージを一時保持するＦＩＦＯ９０１およびＦＩＦＯ９０１に保持されたメッセージをレジスタ１１１１に取り出すＦＩＦＯ読み出し制御１１１０により構成する。ＦＩＦＯ読み出し制御１１１０は、乗り換えスイッチのそれと同様にＦＩＦＯ９０１から出力されるＥＭＰＴＹ信号１１１２が０であることを確認して、リードストローブ１１１３を送出し、ＦＩＦＯ９０１に保持されたメッセージを読みだし、レジスタ１１１１に保持する。このＥＭＰＴＹ信号１１１２はＦＩＦＯ９０１に１個もデータが入ってない時は１で、それ以外は０である。そしてメッセージ受信制御１１０１は、ＦＩＦＯ９０１からメッセージをレジスタ１１１１に取り出したことを、信号１１２２によりメッセージ送出制御１１０２に知らせる。

【００４５】メッセージ送出制御１１０２は、信号１１２２により各メッセージ受信制御１１０１から読みだされたメッセージ内の受信クラスタ番号５０１内のＸ系クロスバ番号／Ｙ系クロスバ番号のどちらを切り出すかを、クロスバスイッチ次元情報８０３により決定する。そして、その受信クラスタ番号５０１から切り出されたＸ／Ｙ系クロスバ番号とポート番号８０４とを一致回路１１２０で比較する。その比較結果が一致すれば、そのポートへメッセージを送出することを決定する。また一致しなければ何も動作はしない。このメッセージの送出では、送出が決定した各メッセージ受信制御１１０１間の送出順序がプライオリティ制御１１２１により決められる。またプライオリティ制御１１２１が送出先に乗り換えスイッチと同様に設けられているＦＩＦＯから出力されるＦＵＬＬ信号９０５が０であることを確認して、ライトストローブ９０５を送出し、送出先のＦＩＦＯにメッセージを書き込む。このＦＵＬＬ信号９０４は送出先ＦＩＦＯが一杯で１個のデータも受け取れない時は１で、それ以外は０である。プライオリティ制御１１２１は送出先のＦＩＦＯにメッセージを書き込んだことを信号１１２３により、メッセージ受信制御１１０１に知らせる。メッセージ受信制御１１０１は、この信号１１２３により次のメッセージを読みだす。またプライオリティ制御１１２１では一回の送出でメッセージの転送が終わらない場合、送り出すごとに信号１１２４により対応するＦＩＦＯ読み出し制御１１１０に対してＦＩＦＯ９０１から次のデータを読みだすことを指示する。

【００４６】次にこのクロスバスイッチ１０１におけるブロードキャストのメッセージ転送の動作について説明する。ＦＩＦＯ読出し制御１１１０がメッセージ内のブロードキャスト制御を判定し、ブロードキャストのメッセージ転送であることを認識すると、ＦＩＦＯ９０１からメッセージをレジスタ１１１１に取り出したことを、信号１１２２により全てのメッセージ送出制御１１０２に知らせる。それぞれのメッセージ送出制御１１０２では、上記と同じ動作を行ない、送出先のＦＩＦＯにメッセージを書き込んだことを信号１１２３により、メッセージ受信制御１１０１に知らせる。メッセージ受信制御１１０１は、信号１１２３が全てのメッセージ送出制御１１０２から送られてくるのを判別する。信号１１２３が全てのメッセージ送出制御１１０２から送られてきたときに、次のメッセージを読みだす。全てから送られてこないときは、全ての信号１１２３が送られてくるまで次のメッセージを読みださない。またプライオリティ制御１１２１では一回の送出でメッセージの転送が終わらない場合、送り出すごとに信号１１２４により対応するＦＩＦＯ読み出し制御１１１０に対してＦＩＦＯ９０１から次のデータを読みだすことを指示する。

【００４７】次にＥＳクラスタのアクセスについて説明する。図１３にＥＳのアドレス空間を示す。ＥＳは図１のようにクロスバスイッチに接続され、分散配置されている。各ソフトウェアからＥＳは各計算クラスタのメモリと独立した１つのアドレス空間にみえる。各ＥＳクラスタ１０５内のＥＳ４０２（図４）は、この１つの空間内の１部の連続領域が割り当てられている。各ＥＳのデータは計算クラスタのメモリに転送してから使用する。計算クラスタから各ＥＳクラスタのアクセスは計算クラスタ間データ転送と同様にクロスバネットワークを介する。各計算クラスタから指定されるＥＳアドレスは、プログラムからＥＳクラスタ間をまたがるアドレスとなる。このため実際のアクセスでは、このアドレスを複数ＥＳクラスタのうち１個を指定するＥＳクラスタ番号と、１個のＥＳクラスタで保持できる最大容量のビット数からなるＥＳクラスタ内アドレスに分けられる。このＥＳクラスタ番号は計算クラスタから指定されるＥＳアドレスの上位の数ビットから変換されて得られる。たとえばＥＳアドレスが３２ビット幅で、各ＥＳクラスタ１０５内のＥＳ４０２（図４）の最大容量が２５６Ｍバイト（2^{28}）であると、ＥＳクラスタ番号に変換されるＥＳアドレスのビット数は上位４ビットである。図１４に、この場合における図１の構成時の変換方法を示す。この場合ＥＳクラスタ１０５は２台だけがクロスバネットワークに接続されているため、ＥＳアドレスの上位４ビットが（００００）$_2$の場合ではＥＳクラスタ番号のＸ系クロスバ番号＝４、Ｙ系クロスバ番号＝２に、またＥＳアドレスの上位４ビットが（０００１）$_2$の場合で

はＥＳクラスタ番号のＸ系クロスバ番号＝４、Ｙ系クロスバ番号＝３に変換される。この変換は、たとえば計算クラスタ１０３内のネットワークインタフェースアダプタ２０２（図２）で実現する。

【００４８】このネットワークインタフェースアダプタ２０２のブロック図を２０図に示す。これはプロセッシングユニット、メモリなどとのインタフェース制御を行なうプロセッサバス制御回路２００１、プロセッサバス制御回路２００１からの指示でメッセージをクロスバネットワークに送出するメッセージ送出制御回路２００２、およびクロスバネットワークから送られてきたメッセージを受信し、クラスタ内のプロセッシングユニット２０１に割込みをかけたり、メッセージで指定された転送データをメモリ２０３に書き込む処理などを制御するメッセージ受信制御回路２００３により構成される。

【００４９】さらにメッセージ送出回路２００２は、メッセージの送出を制御するメッセージ送出制御２２０１、ハードウェアで生成するメッセージの転送データの最大バイト数を保持する最大転送データ長レジスタ２２０２（以下、ＭＴＵと呼ぶ）、プロセッシングユニットから指定される転送データ長を保持する転送データ長レジスタ２２０３（以下、ＬＥＮと呼ぶ）、プロセッシングユニットから指定されるＥＳアドレスを保持するＥＳアドレスレジスタ２２０４（以下、ＥＳＡと呼ぶ）、プロセッシングユニットから指定される自クラスタ内のメモリをアクセスするためのメモリアドレスを保持する送信メモリアドレスレジスタ２２０５（以下、ＭＳ－Ｓと呼ぶ）、プロセッシングユニットから指定される受信側クラスタ内のメモリをアクセスするためのメモリアドレスを保持する受信メモリアドレスレジスタ２２０６（以下、ＭＳ－Ｒと呼ぶ）、プロセッシングユニットから指定される受信クラスタ番号を保持するクラスタ番号レジスタ２２０７（以下、ＣＬＮと呼ぶ）、ＭＴＵ２２０２とＬＥＮ２２０３を比較する比較器２２１１、ＬＥＮ２２０３からＭＴＵ２２０２を減算する減算器２２１２、ＥＳＡ２２０４の上位ビットからＥＳクラスタ番号に変換する変換回路２２１３、ＥＳＡ２２０４とＭＴＵ２２０２を加算する加算器２２１４、ＭＳ－Ｓ２２０４とＭＴＵ２２０２を加算する加算器２２１５、ＭＳ－Ｒ２２０４とＭＴＵ２２０２を加算する加算器２２１６、ＣＬＮ２２０７から受信クラスタ番号に変換する変換回路２２１７、および、メッセージ送出制御の指示により上記レジスタなどの内容を入力とし、メモリから転送データを読出し、メッセージを組立て、クロスバネットワークに送出する送信回路２２０８から構成される。

【００５０】この構成においてＥＳにデータを書き込むための計算クラスタ１０３のメッセージの送出処理について述べる。ＥＳを書き込むためにプロセッシングユニットからは、ネットワークインタフェースアダプタに対して、転送データ長がＬＥＮ２２０３に、ＥＳにデータを書き込むためのＥＳアドレスがＥＳＡ２２０４に、ＥＳに書き込むデータを自クラスタ内のメモリから読みだすためのメモリ主記憶アドレスがＭＳ－Ｓ２２０５に、およびＥＳに対するアクセス種別（書き込み要求）が設定される。ただしアクセス種別についてはメッセージ送出制御２２０１で保持される。またＥＳ書き込みに関係のないレジスタ（この場合ではＭＳ－Ｒ２２０６、ＣＬＮ２２０７である）には何も設定されない。

【００５１】またＭＴＵ２２０２については、プロセッシングユニットから指定された転送データ長がとても長い場合、それをそのままメッセージとして組み立ててしまうと、クロスバネットワークでそのためのメッセージ転送でパスが長い時間、占有されてしまい、他のメッセージ転送に影響がでる。このためメッセージの最大転送データ長を予め決めておき、それよりメッセージが長くなる場合は、その最大転送データ長に従いメッセージを分割して送出する。この最大転送データ長を保持するのがＭＴＵ２２０２であり、この値はシステムが立ち上がる時に設定されている。

【００５２】以下、メッセージ送出回路２００２の動作手順を示す。

【００５３】（１）ＥＳ書き込み処理に必要なレジスタが設定されたことをメッセージ送出制御２２０１が認識する。

【００５４】（２）比較器２２１１においてプロセッシングユニットから指定された転送データ長ＬＥＮ２２０３がハードウェアで生成するメッセージの最大転送データ長ＭＴＵ２２０３以下であるかを判定し、その結果をメッセージ送出制御２２０１に渡す。

【００５５】（３）メッセージ送出制御２２０１では、プロセッシングユニットから指定された転送データ長がハードウェアで生成するメッセージの最大転送データ長を越えている場合には、選択回路２２０９を制御し、転送データ長をＭＴＵ２２０３を選択し、送信回路２２０８に送る。またＥＳＡ２２０４の上位のあるビット幅分は変換回路２２１３によりＥＳクラスタ番号に変換され、送信回路２２０８に送られる。ＥＳＡ２２０４の残りのビットおよびＭＳ－Ｓ２２０５はそのまま送信回路２２０８に送られる。

【００５６】（４）そしてメッセージ送出制御２２０１は、送信回路２２０８に対してメッセージの送信を行なうことを指示する。

【００５７】（５）送信回路２２０８では、ＭＳ－Ｓ２２０６に従い、クラスタ内のメモリから転送データを転送データ長分を読出し、それとメッセージ転送に必要な情報を組み合わせてメッセージとしてクロスバネットワークに送出する。

【００５８】（６）送信回路２２０８がメッセージの送出を終えると、メッセージ送出制御２２０１に知らせる。

【００５９】（７）メッセージ送出制御２２０１では送信回路２２０８からのメッセージの送出終了を知ると、ＥＳＡ２２０４およびＭＳ－Ｓ２２０５の値にＭＴＵ２２０２の値をそれぞれの加算器２２１４、２２１５にて加算を行なう。この加算結果はそれぞれのＥＳＡ２２０４およびＭＳ－Ｓ２２０５に設定する。

【００６０】（８）この後、減算器２２１２にてＬＥＮ２２０３からＭＴＵ２２０２を減算し、その結果をＬＥＮ２２０３に設定する。

【００６１】（９）（２）から（８）までをＬＥＮ２２０３で保持する値がＭＴＵ２２０２で保持する値以下になるまで繰り返す。

【００６２】（１０）ＬＥＮ２２０３で保持する値がＭＴＵ２２０２で保持する値以下になると、メッセージ送出制御２２０１では選択回路２２０９を制御し、転送データ長をＬＥＮ２２０３を選択し、送信回路に送り、（２）から（８）までの処理を１回だけ行なうことにより、ＥＳにデータを書き込むための計算クラスタ１０３のメッセージの送出処理を終える。

【００６３】ここでＥＳのアクセスが複数のＥＳクラスタに渡る場合にでも、たとえばＥＳのアクセスを４Ｋバイト単位で、かつ４Ｋバイト境界とし、またハードウェアで生成するメッセージの最大転送データ長ＭＴＵ２２０３を８Ｋバイトにすれば、このＥＳのアドレス加算の結果を番号変換することにより、複数のＥＳクラスタに渡っても、この構成で処理できる。

【００６４】また、このネットワークインタフェースアダプタ２０２は、計算クラスタ１０３、入出力クラスタ１０４、ＥＳクラスタ１０５、ＦＥＰ１０６の中に構成され、計算クラスタ間のデータ転送でも、ＭＴＵ２２０１、ＬＥＮ２２０３、ＭＳ－Ｓ２２０４、ＭＳ－Ｒ２２０６、ＣＬＮ２２０７などを用いてＥＳのアクセスと同様に実現できる。

【００６５】図１５にＥＳアクセスのメッセージフォーマットを示す。このメッセージには、大きく計算クラスタ間での排他制御を行なうためのロック処理、読み出し処理、および書き込み処理がある。これらのメッセージでは計算クラスタ間データ転送メッセージと同様に受信クラスタ番号、転送制御および、それぞれの処理に対応したデータ転送コマンドが付けられる。

【００６６】ロック処理では、ＥＳ上にロック変数を定義し、それにより計算クラスタ間で排他制御を行なう。まず計算クラスタからのロック要求メッセージ（ａ）がＥＳクラスタ内のネットワークインタフェースアダプタ２０２に届くと、そのメッセージをＥＳ制御プロセッサ４０１に渡す。ＥＳ制御プロセッサ４０１では、そのメッセージ内のＥＳアドレスに従い、ＥＳ４０２からＥＳデータを読み出す。そして、そのＥＳデータとメッセージ内の比較データとを比較し、等しければメッセージ内の格納データをそのＥＳアドレスに格納する。等しくなければ格納はしない。この比較、格納する処理を行なう間は、他のＥＳアクセスに関係する処理はネットワークインタフェースアダプタ２０２で行なわないようにする。またＥＳ制御プロセッサ４０１では、その比較結果に従い条件コードを作成する。そしてロック応答メッセージ（ｂ）を計算クラスタに送り返すためにネットワークインタフェースアダプタ２０２に必要な情報を設定し、依頼する。ネットワークインタフェースアダプタ２０２では、ロック応答メッセージ（ｂ）を組立て、ネットワークにメッセージを送出する。計算クラスタに受信されたロック応答メッセージ（ｂ）は、その中のプロセッシングユニット２０１（図２）に渡され、その条件コードの内容によりロック処理が行なわれたことを知る。

【００６７】またロックを解除するアンロック処理では計算クラスタからアンロック要求メッセージ（ｃ）を送出する。ＥＳクラスタでそのメッセージが届くと、ＥＳ制御プロセッサ４０１にそのメッセージを渡すことなく、ネットワークインタフェースアダプタ２０２がそのメッセージ内のＥＳアドレスに従い格納データを書き込む。これによりアンロック処理は終了する。

【００６８】ＥＳの読み出し処理では、計算クラスタは前述したＥＳアドレスの変換を行ない、読み出し要求メッセージ（ｄ）を送出する。読み出し要求メッセージ（ｄ）がいずれかのＥＳクラスタ内のネットワークインタフェースアダプタ２０２（図４）に届くと、そのメッセージをＥＳ制御プロセッサ４０１に渡す。ＥＳ制御プロセッサ４０１ではそのメッセージを解釈し、ＥＳデータの読出しに必要な情報をネットワークインタフェースアダプタ２０２に設定し、ＥＳデータ読み出し応答メッセージ（ｅ）の送出を依頼する。ネットワークインタフェースアダプタ２０２では、ＥＳ制御プロセッサ４０１から設定された情報に従い、読み出しデータ長分だけＥＳデータを読みだし、読み出し応答メッセージ（ｅ）を作成し、ネットワークに送出する。読み出し応答メッセージ（ｅ）を受け取った計算クラスタでは、ネットワークインタフェースアダプタ２０２（図２）がそのメッセージ内の読み出しデータ格納アドレスに従いメモリ２０３（図２）に書き込み、ＥＳ読み出しデータが届いたことをプロセッシングユニット３０１（図２）に知らせる。

【００６９】ＥＳの書き込み処理では、計算クラスタは前述したＥＳアドレスの変換を行ない、書き込み要求メッセージ（ｆ）を送出する。そのメッセージを受信したＥＳクラスタ１０５内のネットワークインタフェースアダプタ２０２では、メッセージ内のＥＳアドレスに従い、書き込みデータ長分だけＥＳ４０２に書き込みデータを書き込む。

【００７０】次に入出力について説明する。入出力クラスタは図１のようにＸ方向クロスバスイッチ１０１ｘとＹ方向クロスバスイッチ１０１ｙに接続され、分散配置

されている。各入出力クラスタのアクセスは計算クラスタ間データ転送と同様にクロスバネットワークを介してメッセージ転送により行なわれる。入出力クラスタがＸ方向クロスバスイッチ１０１ｘとＹ方向クロスバスイッチ１０１ｙに接続されている理由は、全計算クラスタ１０３から各入出力クラスタ１０４に対してアクセスが可能にするためである。もし仮にＸ方向のクロスバスイッチに接続されている計算クラスタが同じクロスバスイッチに接続されている入出力クラスタだけにしかアクセスしないのであればＹ方向のクロスバスイッチはなくてもよい。

【００７１】計算クラスタ１０３から入出力を行う場合は入出力命令により実行される。この入出力命令では、入出力クラスタ番号、入出力装置番号あるいはファイル名などが指定され、計算クラスのネットワークインタフェースアダプタに渡される。このネットワークインタフェースアダプタでは、計算クラスタ間データ転送メッセージと同様に命令で指定された入出力クラスタ番号からハードウェアで使用する受信クラスタ番号に変換を行なう。またメッセージのハードウェアヘッダとして転送制御、転送データ長などを生成する。メッセージ内の転送データは、入出力装置に書き込むためのデータや入出力のプロトコルなどとして使用する情報であり、それとハードウェアヘッダをメッセージとしてクロスバネットワークに送出する。

【００７２】メッセージを受信した入出力クラスタのネットワークインタフェースアダプタ２０２（図３）は入出力クラスタの入出力制御プロセッサ３０２にその制御を渡す。入出力制御プロセッサではメッセージの転送データの内容を解読し、入出力アダプタ２０４、ＬＡＮアダプタ３０２などを介してメッセージの転送データの内容に対応して実際の入出力装置との間で処理を行なう。この処理についてはＵＮＩＸオペレーティングシステムなどの処理と同様でよい。また入出力動作後、計算クラスタ１０３への応答が必要な場合は、入出力制御プロセッサ３０２がネットワークインタフェースアダプタ２０２にその処理を依頼する。

【００７３】次に本発明による並列プロセッサの実装について説明する。図１６に図１の構成時の実装を示す。１６０１はまとめて実装する１つの単位であり、たとえば筐体、パッケージ、ＬＳＩなどのいずれかに対応する。この実装単位１６０１は１個のＹ方向のクロスバスイッチ１０１ｙとそれに接続されるそれぞれ４つの乗り換えスイッチ１０２と計算クラスタ１０３、および１個のＸ方向のクロスバスイッチ１０１ｘである。

【００７４】また入出力クラスタ１０４とＥＳクラスタ１０５の実装については、実装単位１６０１内の計算クラスタ１０３を入出力クラスタ１０４とＥＳクラスタ１０５に置き換えて実装する。さらにＦＥＰ１０６とそれに接続する１個のブロードキャストスイッチ１００は、上記とは別に実装する。

【００７５】この実装単位１６０１を４つと、その単位１６０１内の計算クラスタ１０３を入出力クラスタ１０４とＥＳクラスタ１０５に置き換えた単位を１つと、ＦＥＰ１０６とそれに接続する１個のブロードキャストスイッチ１００の単位により、図１の構成の並列プロセッサが実現される。

【００７６】この実装ではネットワークの構成要素であるクロスバスイッチ１０１を分散実装が可能である。これによりクロスバスイッチ群とクラスタ群間を接続するインタフェースが１ヶ所に集中することがない。またクロスバスイッチ群とクラスタ群に分けて、それぞれを集中実装した場合と比較し、乗り換えスイッチとクロスバスイッチとの接続するインタフェースにおいて、実装単位間を接続するインタフェースの本数を減らすことができる。たとえば実装単位１６０１が筐体であるならば筐体間を渡るケーブルの本数を減らすことができる。またクラスタの台数を拡張する場合にも、予めクロスバスイッチ１０１のポート数を増やしておけば、その実装単位１６０１で増設が可能である。さらにこの場合、実装単位１６０１内でクラスタとクロスバスイッチのインタフェースが閉じるため、その間のインタフェースは近距離になり、実装単位を渡るインタフェースに比べ、高速なクラスタ間データ転送が実現できる。この局所的に速いクラスタ間データ転送を考慮し、並列処理のクラスタへの割当てなどを行なえば、並列処理全体の性能を向上することも可能である。

【００７７】また乗り換えスイッチ、各クラスタ、クロスバスイッチ間を接続するインタフェースは高速に実現可能な同期転送を用いることが考えられる。しかし、この同期転送ではクラスタの最大台数時のインタフェースによるディレイを考慮して設計を行なう必要があり、またクラスタの台数を拡張するごとにクロックスキュー合わせが必要になる。これを考えれば乗り換えスイッチ、各クラスタ、クロスバスイッチ間を接続するインタフェースは非同期転送にするのが好ましい。このため本実施例では前述したように、各クラスタと乗り換えスイッチ間、乗り換えスイッチとクロスバスイッチ間にＦＩＦＯを設け、非同期転送を実現する。

【００７８】ＦＩＦＯの動作は図１７に示すように、書き込み動作ではＦＩＦＯから出力されるＦＵＬＬ信号が０であることを確認して、ライトストローブとＦＩＦＯ書き込みデータを送出し、ＦＩＦＯにそのデータを書き込む。このＦＵＬＬ信号は送出先ＦＩＦＯが一杯で１個のデータも受け取れない時は１で、それ以外は０である。またＦＵＬＬ信号が１である時は、ＦＵＬＬ信号が０になる、即ち読み出し動作が行なわれるまで書き込み動作を抑止する。一方、読み出し動作ではＦＩＦＯから出力されるＥＭＰＴＹ信号が０であることを確認して、リードストローブを送出し、ＦＩＦＯに保持されたメッ

セージを読みだす。このＥＭＰＴＹ信号はＦＩＦＯに１個もデータが入ってない時は１で、それ以外は０である。またＥＭＰＴＹ信号が１である時は、ＥＭＰＴＹ信号が０になる、即ち書き込み動作が行なわれるまで読み出し動作を抑止する。また、これらの書き込み動作と読み出し動作は上記の動作を守れば、非同期に行なうことができる。

【００７９】これを用いれば各クラスタ、乗り換えスイッチ、クロスバスイッチ間をそれぞれ独立に動作することが可能になる。これにより、たとえば乗り換えスイッチにエラーが生じ、その解析のためサービスプロセッサ（以下、ＳＶＰ）がその乗り換えスイッチに対して乗り換えスイッチ内のレジスタ群などをスキャンアウト動作をしている時も、それに接続される各クロスバスイッチやクラスタは、そのスキャンアウト動作を知る必要がなく、また、それにより停止する必要もなく通常動作を実行できる。これは、各クロスバスイッチやクラスタの動作が、乗り換えスイッチに対するデータ転送動作を非同期転送であるＦＩＦＯの書き込み／読み出し動作を用いており、そのＦＩＦＯから出力されるＦＵＬＬ／ＥＭＰＴＹ信号により、抑止できるためである。このように乗り換えスイッチ、各クラスタ、クロスバスイッチ間を接続するインタフェースを非同期転送にすることにより、それらを分離して部分保守が可能になる。

【００８０】以上、本実施例では１６台の計算クラスタが２次元アレイ状に配置された場合について説明したが、各クラスタの台数、その次元数を限定する必要はなく同様な考え方で実施可能である。図１８に３次元アレイ状に配置された並列プロセッサの構成のイメージを示す。ここで、１８０１は乗り換えスイッチ１０２と計算クラスタ１０３を、１８０２は乗り換えスイッチ１０２と入出力クラスタ１０４を、１８０３は乗り換えスイッチ１０２とＥＳクラスタ１０５を、１つにしてそれぞれ図示している。この詳細な説明は省略するが、このように容易に実現できることがわかる。

【００８１】またＦＥＰ１０６の接続においても新たなクロスバスイッチ１００を設けるのではなく、図１９に示すように図１の入出力クラスタ１０４、ＥＳクラスタ１０５などの代わりに複数ある構成であってもよい。この構成ではＦＥＰ１０６の負荷が重く、１台で処理しきれない場合に、複数のＦＥＰ１０６を接続し、その負荷を分散するのに有効である。

【００８２】なお以上の実施例においては、Ｘ方向のクロスバスイッチとＹ方向のクロスバスイッチとＦＥＰに接続するクロスバネットワークを全て同一構造と仮定した。従って、それらのポートの数は全て同一とした。しかし本発明はこれらに限られるのでなく、Ｘ方向のクロスバスイッチとＹ方向のクロスバスイッチに接続するプロセッサの数は異なっていてもよく、従って、それらのポートの数は異なっていてもよい。またクロスバネットワークは、ＦＥＰから全あるいは各クラスタへの転送および各クラスタからＦＥＰへの転送が行える必要最小限の機能があればよい。

【００８３】

【発明の効果】本発明によれば、２のべき乗＋１個のポートを持つクロスバスイッチを設け、そのクロスバスイッチに２のべき乗個となるように実行プロセッサ群を配置し、クロスバスイッチの残りの１ポートに並列処理の管理機能や入出力機能を実行する補助プロセッサを配置することができるので、各実行プロセッサにポートの追加、補助プロセッサを結合するためのクロスバスイッチの追加は必要なくなり、並列プロセッサの実装規模を小さくできる。さらに並列処理の実行は並列処理の管理機能や入出力機能とは独立に実行でき、並列処理を実行するプロセッサの台数を減らすことなく２のべき乗台とし、並列プロセッサを構成できる。

【００８４】また本発明では、ある１つの次元のクロスバスイッチ、そのクロスバスイッチに接続するプロセッサ群、および、そのプロセッサ群の中の１個のプロセッサに接続する全ての、上記とは別の次元のクロスバスイッチを、１つの実装単位とするため、並列プロセッサの実装においては、クロスバスイッチ群を実装するための専用のＬＳＩ、筐体などは必要なく、この実装単位だけの組合せで可能である。これによりプロセッサとネットワークを接続するインタフェースを一ヶ所に集中させることなく実装できる。

【図面の簡単な説明】

【図１】本発明による２次元構成時の並列プロセッサの一実施例を示すシステム構成図。

【図２】図１の構成に用いる計算クラスタの概略構成図。

【図３】図１の構成に用いる入出力クラスタの概略構成図。

【図４】図１の構成に用いる拡張記憶クラスタの概略構成図。

【図５】本発明による計算クラスタ間データ転送メッセージフォーマット図。

【図６】本発明によるデータ転送メッセージ内の受信クラスタ番号を示す図。

【図７】本発明によるデータ転送メッセージ内の転送制御を示す図。

【図８】図１の構成に用いるデータ転送のルーティング情報を示す図。

【図９】図１の構成に用いる乗り換えスイッチの概略構成図。

【図１０】図９の構成に用いる転送先決定回路を示す図。

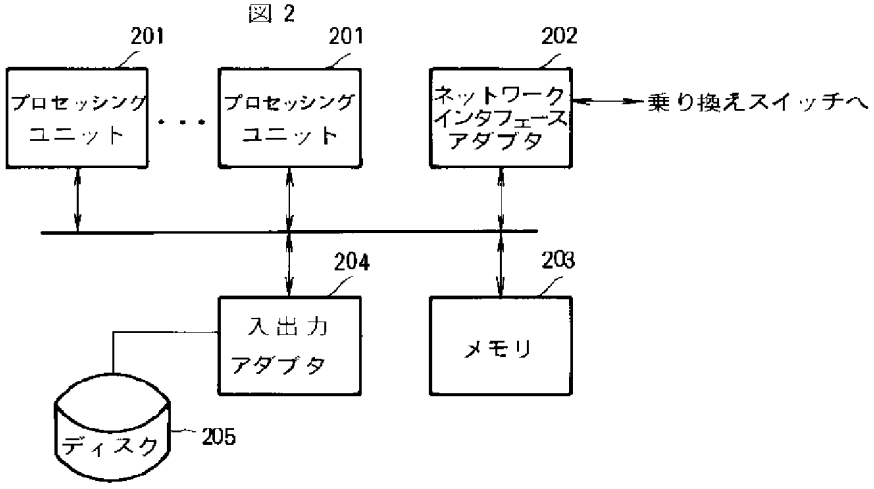
【図１１】図１の構成に用いるクロスバスイッチの概略構成図。

【図１２】図１の構成に用いる計算クラスタ番号変換を

示す図。
【図１３】本発明による拡張記憶の空間を示す図。
【図１４】図１の構成に用いる拡張記憶クラスタ番号変換を示す図。
【図１５】本発明による拡張記憶アクセスのメッセージフォーマット図。
【図１６】図１の構成における並列プロセッサの実装を示す図。
【図１７】本発明で用いるＦＩＦＯのタイムチャート図。
【図１８】本発明による３次元構成時の並列プロセッサの一実施例を示すシステム構成イメージ図。
【図１９】本発明によるフロントエンドを複数結合した２次元構成時の並列プロセッサの一実施例を示すシステム構成イメージ図。
【図２０】本発明によるネットワークインタフェースアダプタのブロック図。
【符号の説明】
１０１…クロスバスイッチ、１０２…乗り換えスイッチ、１０３…計算クラスタ、１０４…入出力クラスタ、１０５…拡張記憶クラスタ、１０６…フロントエンド、１０７…ワークステーション、１０８…ファイルサーバ、１０９…ディスク装置、１１０…ローカルエリアネットワーク（ＬＡＮ）、１１１…ディスク装置、２０１…プロセッシングユニット、２０２…ネットワークインタフェースアダプタ、２０３…メモリ、２０４…入出力アダプタ、２０５…ディスク装置、３０１…入出力制御ユニット、３０２…ＬＡＮアダプタ、９０１…ファーストインファーストアウトメモリ（ＦＩＦＯ）、９０４…ＦＩＦＯから出力されるＦＵＬＬ信号、９１１…ＦＩＦＯから出力されるＥＭＰＴＹ信号、１６０１…まとめて実装する１つの単位。

【図２】

図 2

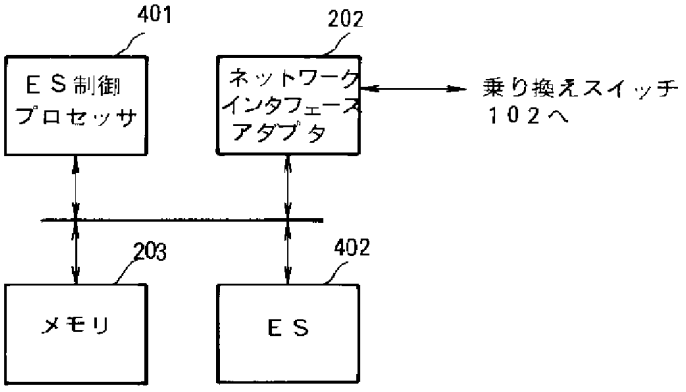


【図６】

図 6

【図４】

図 4

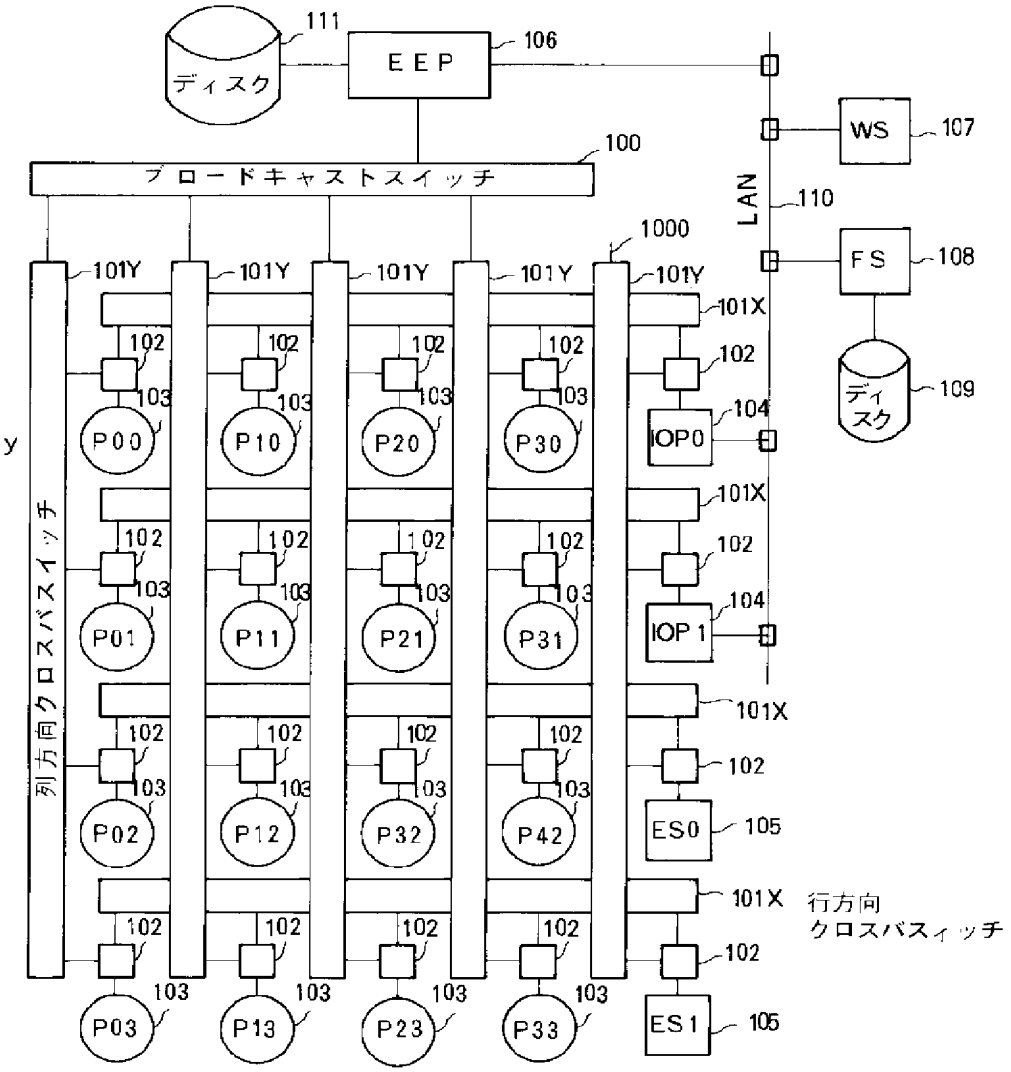


【図７】

図 7

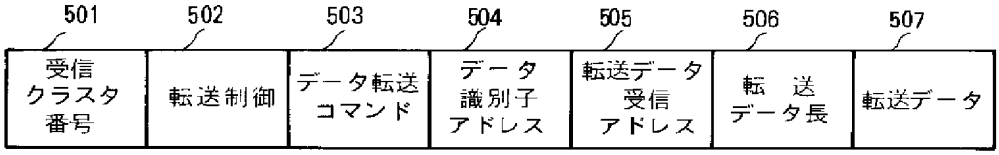
【図1】

図 1


111
ディスク
EEP
106
100
ブロードキャストスイッチ
WS
107
LAN
110
FS
108
ディスク
109
1000
101Y
101X
102
103
104
P00
P10
P20
P30
IOP0
P01
P11
P21
P31
IOP1
P02
P12
P32
P42
ES0
105
P03
P13
P23
P33
ES1
y
列方向クロスバスイッチ
行方向
クロスバスイッチ

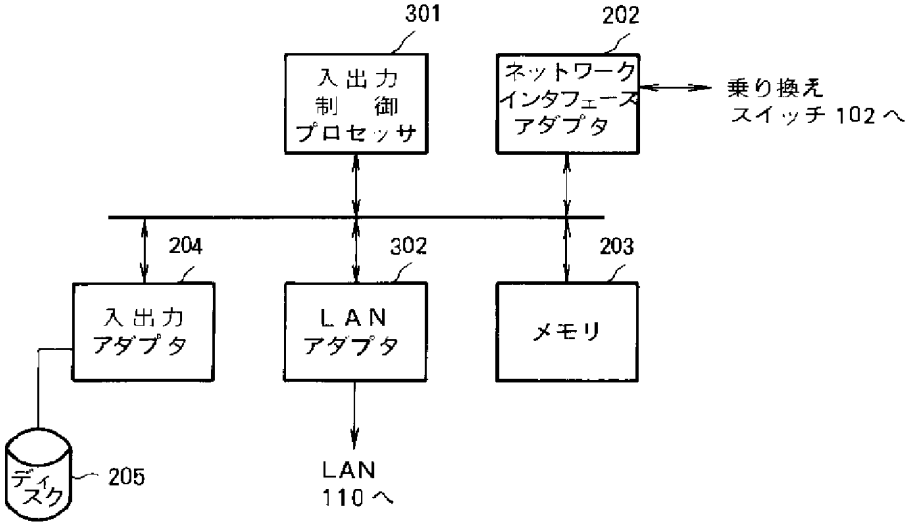
【図5】

図 5


501
502
503
504
505
506
507
受信
クラスタ
番号
転送制御
データ転送
コマンド
データ
識別子
アドレス
転送データ
受信
アドレス
転　送
データ長
転送データ

【図３】

図 3



【図１４】

図 14

入力	出力 X	出力 Y
0 0 0 0	4	2
0 0 0 1	4	3

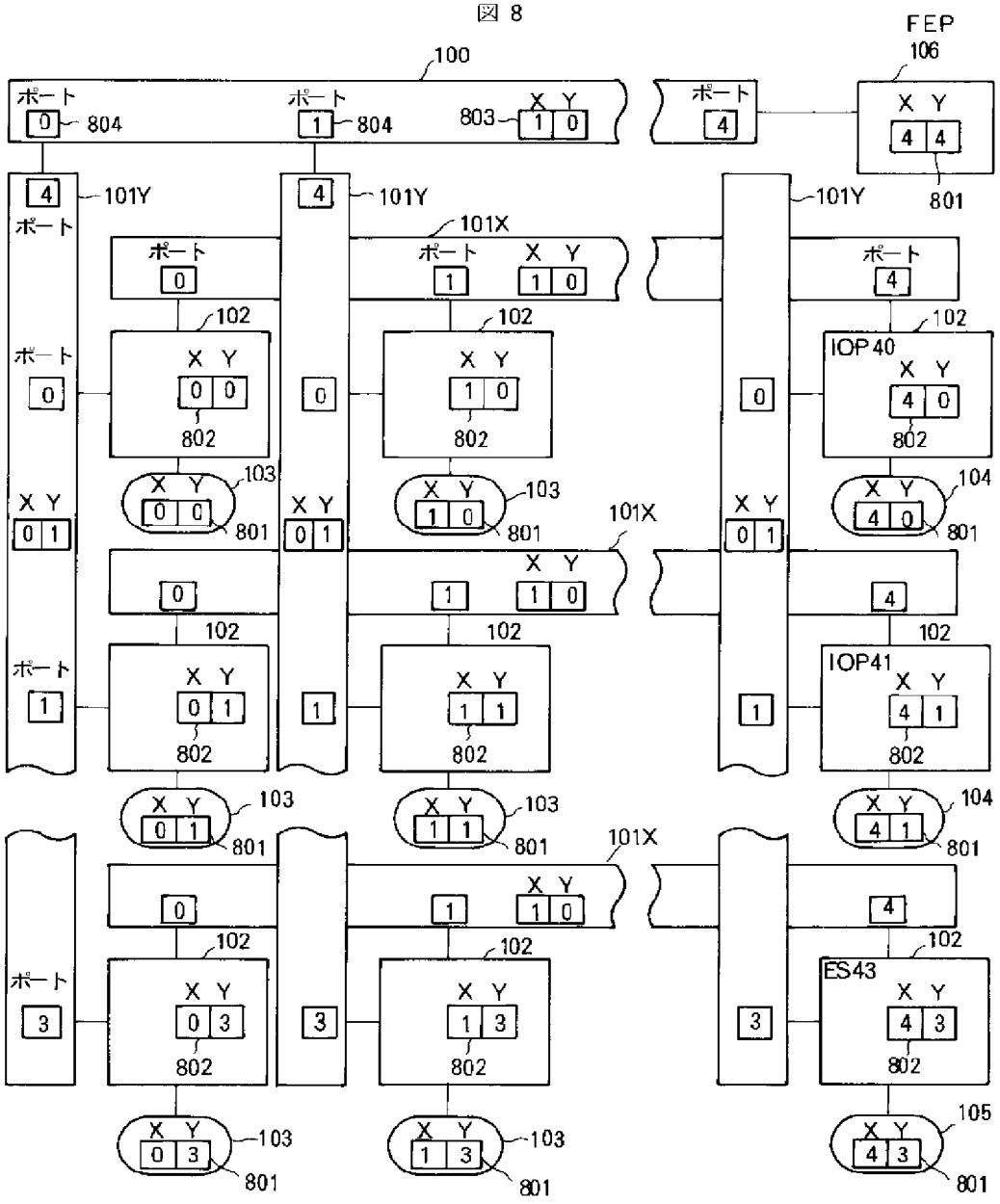
【図１０】

受信クラスタ番号の比較結果 X系	受信クラスタ番号の比較結果 Y系	転送順序制御 0（X系クロスバから転送）	転送順序制御 1（Y系クロスバから転送）
0（不一致）	0（不一致）	X系クロスバに転送	Y系クロスバに転送
0	1（一致）	同　上	同　上
1（一致）	0	Y系クロスバに転送	X系クロスバに転送
1	1	クラスタに転送	クラスタに転送

図 10

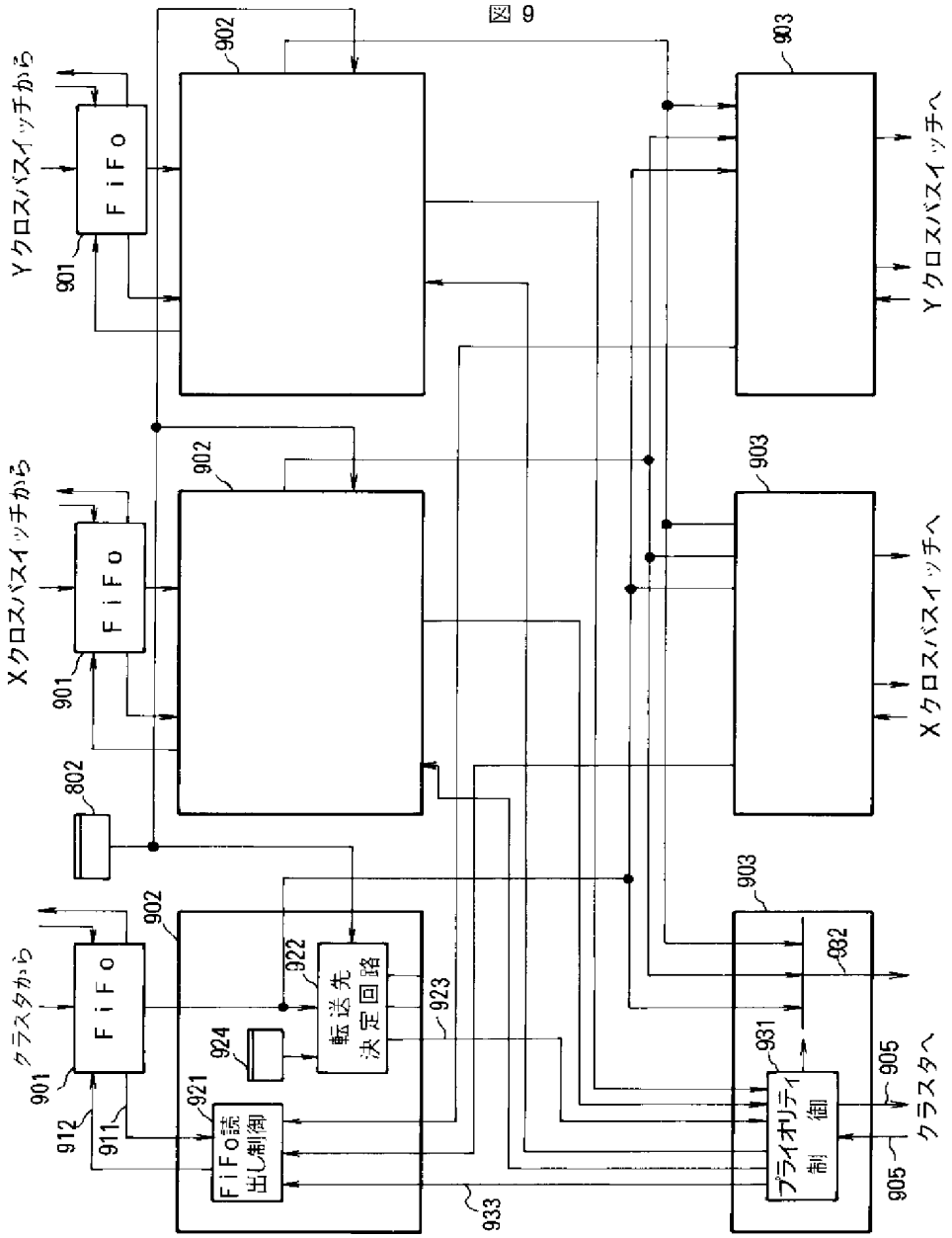
【図８】

図 8


FEP
106
100
ポート
0
804
ポート
1
804
803
X Y
1 0
ポート
4
X Y
4 4
801
4
101Y
ポート
101X
ポート
0
ポート
1
X Y
1 0
ポート
4
102
ポート
0
X Y
0 0
802
X Y
0 1
103
801
0
1
X Y
1 0
X Y
4 0
IOP40
104
ポート
1
X Y
0 1
IOP41
X Y
4 1
ポート
3
X Y
0 3
ES43
X Y
4 3
105

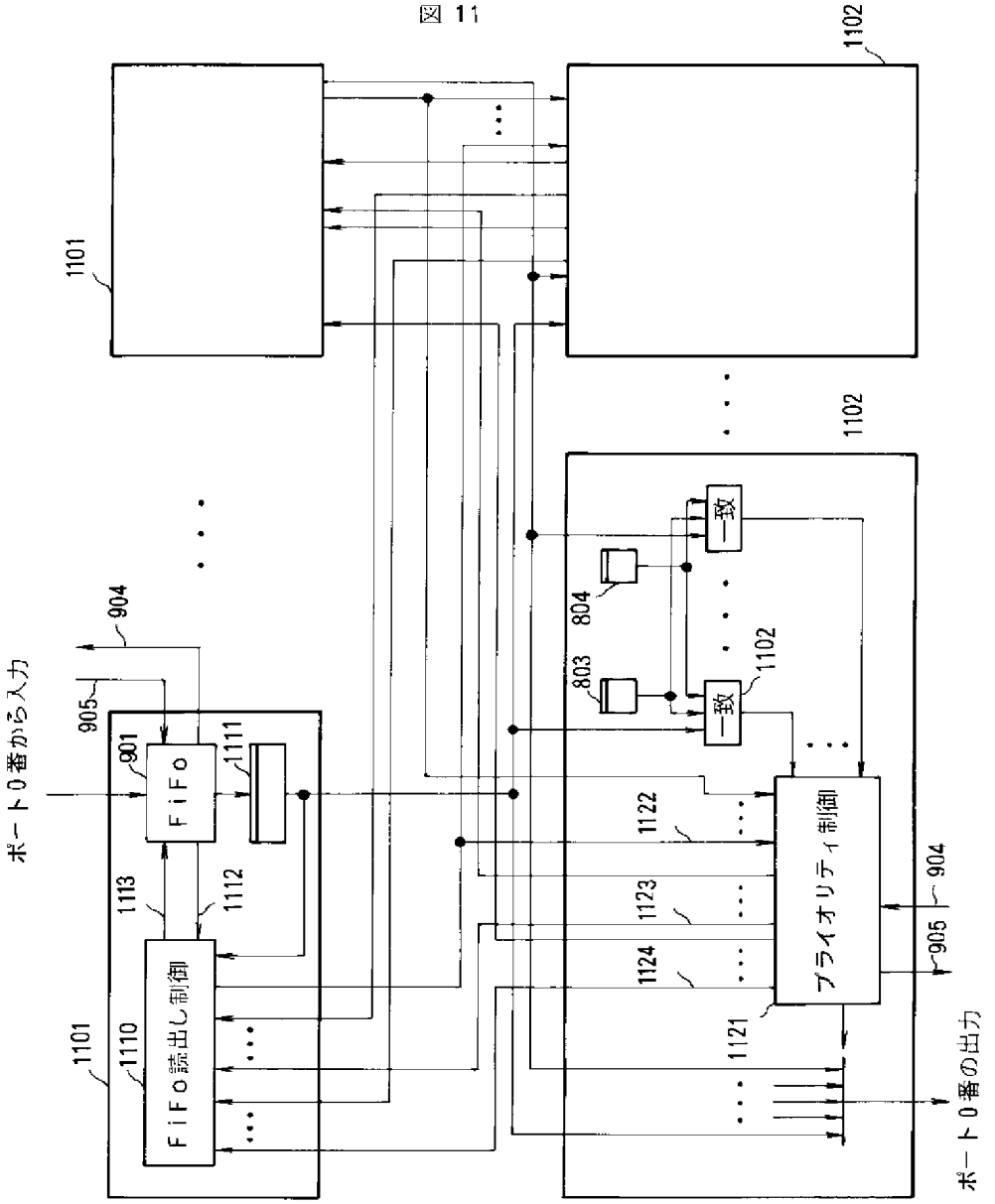
【図9】

図 9


Yクロスバスイッチから
Xクロスバスイッチから
クラスタから
FiFo
901
902
903
802
912
911
921
FiFo読出し制御
922
転送先決定回路
923
924
931
932
プライオリティ制御
905
933
Yクロスバスイッチへ
Xクロスバスイッチへ
クラスタへ

【図11】

図 11


1101
1102
1102
1102
ポート0番から入力
904
905
901
FiFo
1111
1112
1113
FiFo読出し制御
1110
1101
804
803
一致
一致
1122
1123
1124
1121
プライオリティ制御
904
905
ポート0番の出力

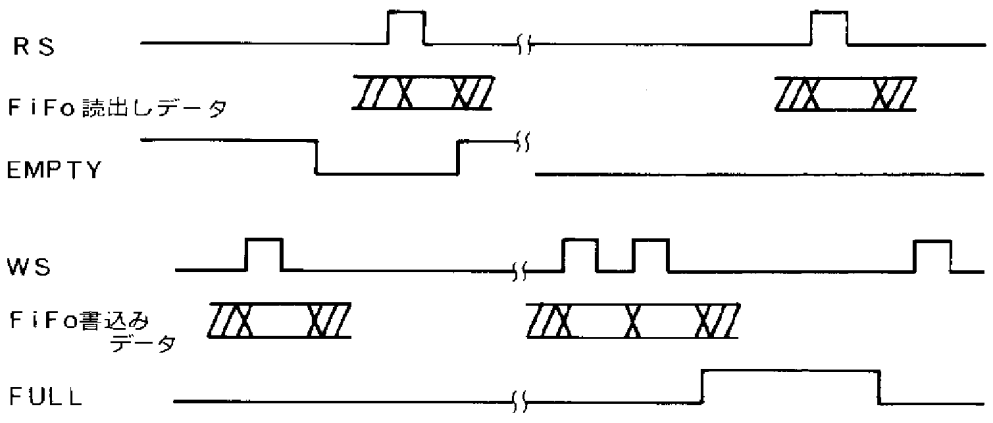
【図１２】

図 12

入　力		出　力	
意　味	ビットパターン	X	Y
P 0 0	0 0 0 0 0	0 0 0	0 0 0
P 1 0	0 0 0 0 1	0 0 1	0 0 0
P 2 0	0 0 0 1 0	0 1 0	0 0 0
P 3 0	0 0 0 1 1	0 1 1	0 0 0
P 0 1	0 0 1 0 0	0 0 0	0 0 1
P 1 1	0 0 1 0 1	0 0 1	0 0 1
P 2 1	0 0 1 1 0	0 1 0	0 0 1
P 3 1	0 0 1 1 1	0 1 1	0 0 1
P 0 2	0 1 0 0 0	0 0 0	0 1 0
P 1 2	0 1 0 0 1	0 0 1	0 1 0
P 2 2	0 1 0 1 0	0 1 0	0 1 0
P 3 2	0 1 0 1 1	0 1 1	0 1 0
P 0 3	0 1 1 0 0	0 0 0	0 1 1
P 1 3	0 1 1 0 1	0 0 1	0 1 1
P 2 3	0 1 1 1 0	0 1 0	0 1 1
P 3 3	0 1 1 1 1	0 1 1	0 1 1
IOP0	1 0 0 0 0	1 0 0	0 0 0
IOP1	1 0 0 0 1	1 0 0	0 0 1
F E S	1 0 0 1 0	1 0 0	1 0 0

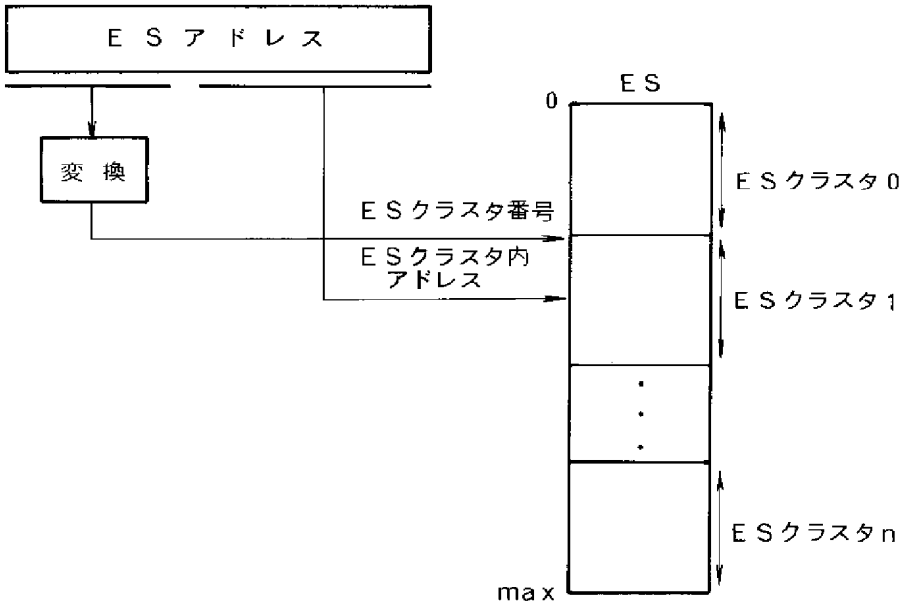
【図１７】

図 17



【図１３】

図 13


ＥＳアドレス
変換
ＥＳクラスタ番号
ＥＳクラスタ内
アドレス
ES
0
max
ＥＳクラスタ０
ＥＳクラスタ１
ＥＳクラスタｎ

【図１５】

図 15

(a) ロック要求

受信（ＥＳ）クラスタ番号	転送制御	ロック要求	送信クラスタ番号	ＥＳアドレス	比較データ	格納データ

(b) ロック応答

受信クラスタ番号	転送制御	ロック応答	条件コード

(c) アンロック要求

受信（ＥＳ）クラスタ番号	転送制御	アンロック要求	ＥＳアドレス	格納データ

(d) 読出し要求

受信（ＥＳ）クラスタ番号	転送制御	読出し要求	送信クラスタ番号	ＥＳアドレス	読出しデータ格納アドレス	読出しデータ長

(e) 読出し応答

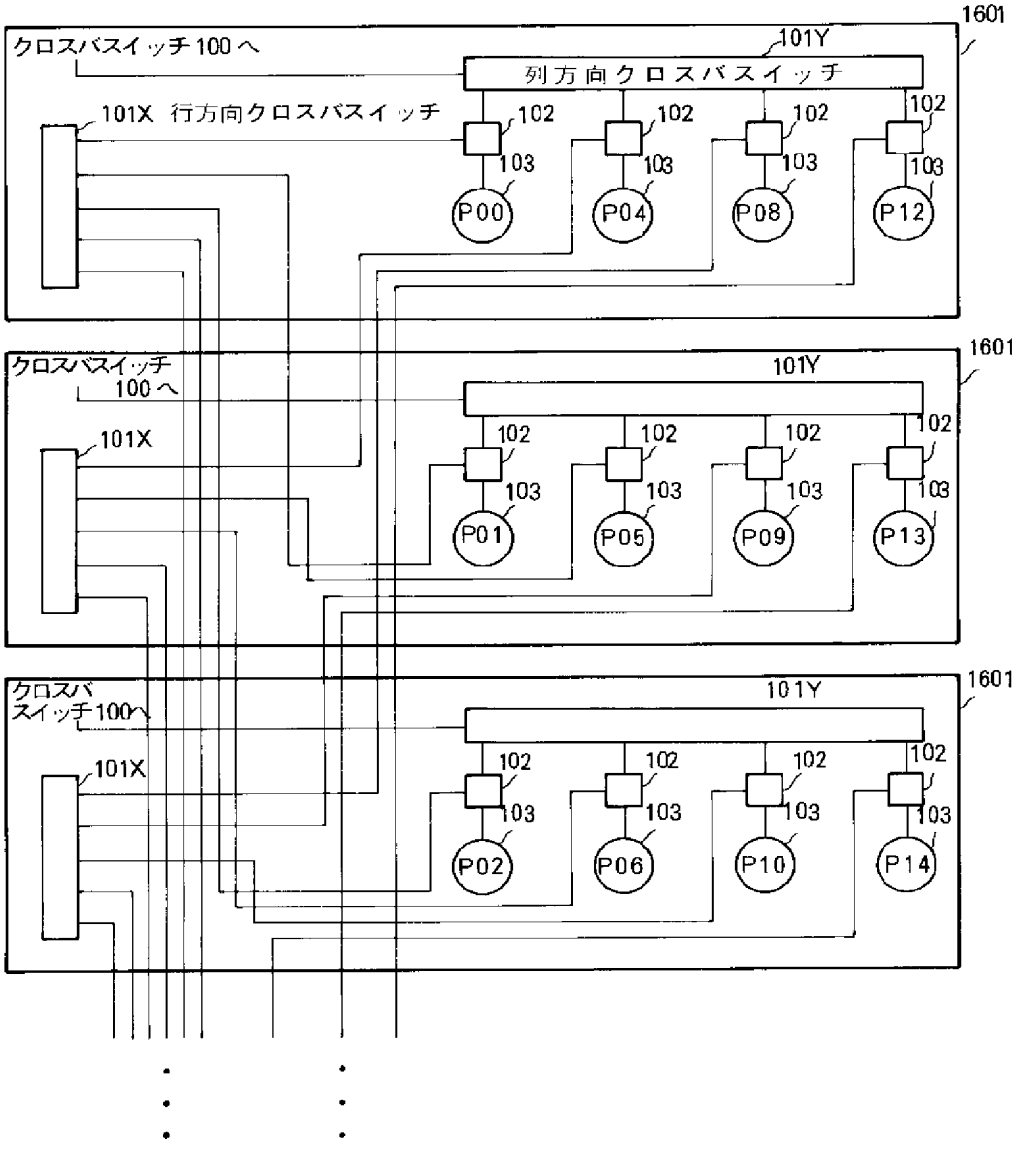
受信クラスタ番号	転送制御	読出し応答	読出しデータ格納アドレス	読出しデータ長	読出しデータ

(f) 書込み要求

受信（ＥＳ）クラスタ番号	転送制御	書込み要求	ＥＳアドレス	書込みデータ長	書込みデータ

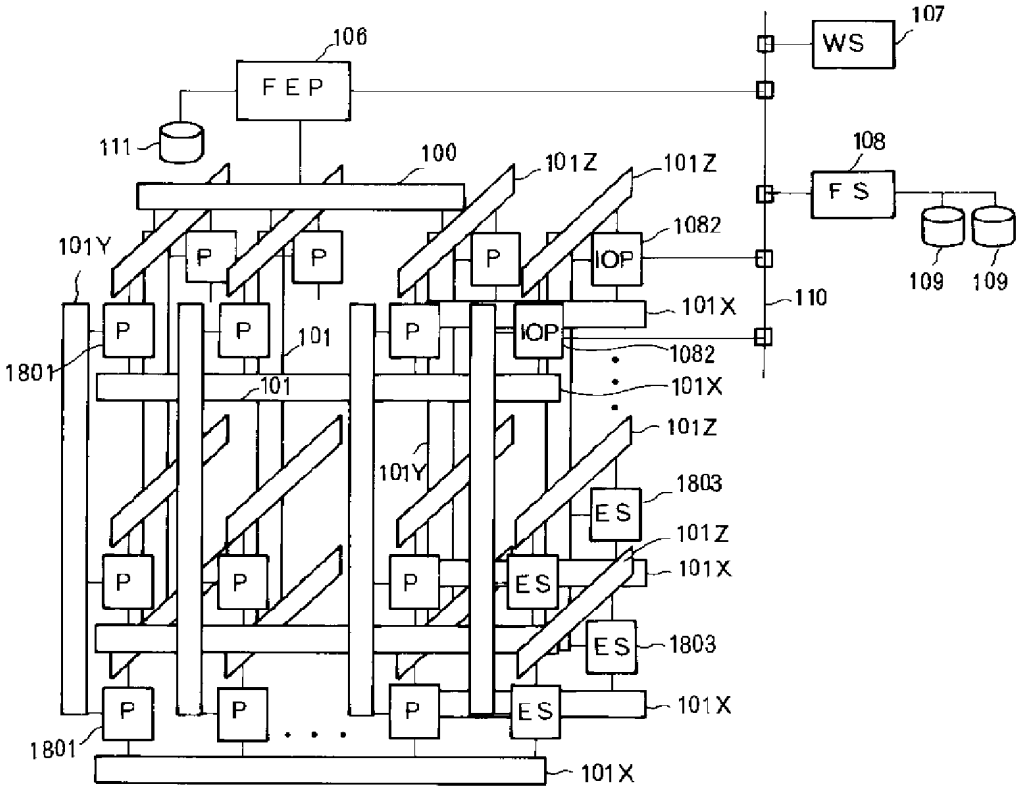
【図１６】

図 16


1601
クロスバスイッチ100へ
101Y
列方向クロスバスイッチ
101X 行方向クロスバスイッチ
102
103
P00
P04
P08
P12
1601
クロスバスイッチ
100へ
101Y
101X
P01
P05
P09
P13
1601
クロスバ
スイッチ100へ
101Y
101X
P02
P06
P10
P14

【図１８】

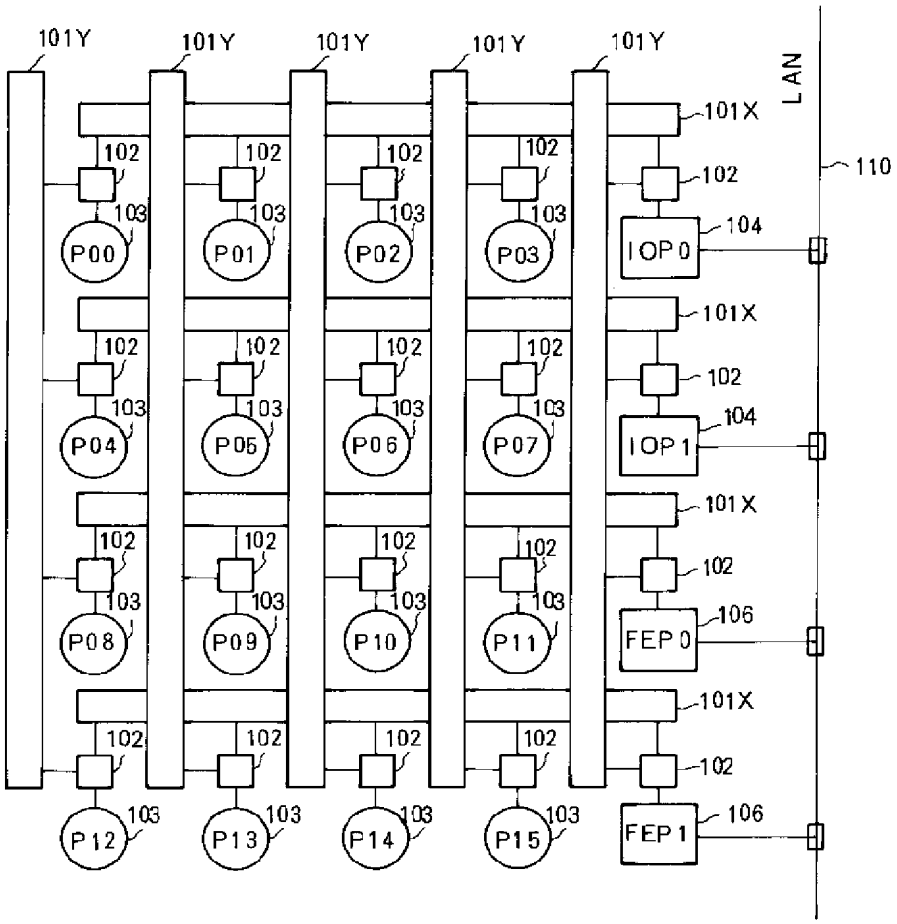
図 18



１８０１：計算クラスタと乗り換えスイッチ
１８０２：入出力クラスタと乗り換えスイッチ
１８０３：ＥＳクラスタと乗り換えスイッチ

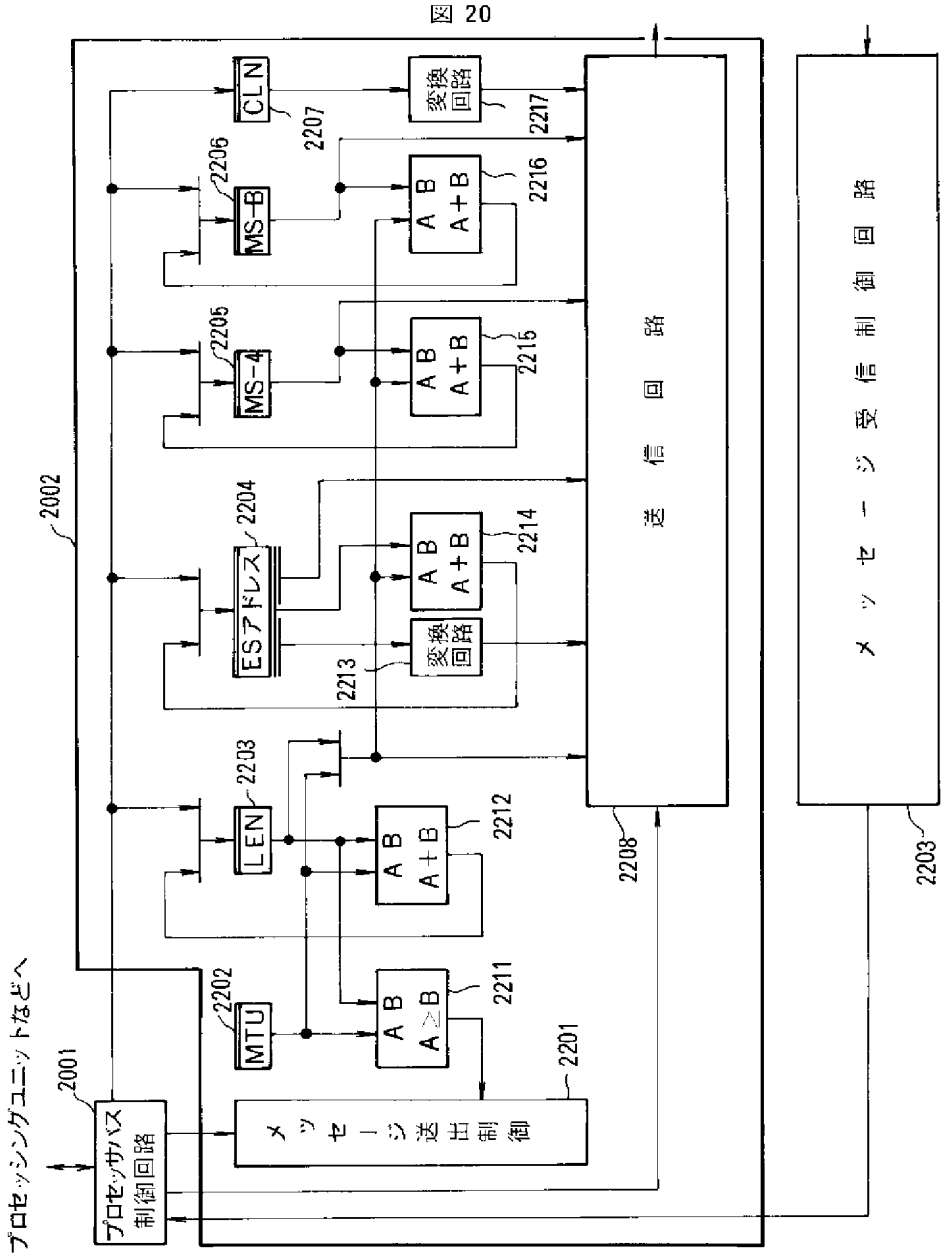
【図１９】

図 19


101Y
101X
LAN
110
102
103
104
106
P00
P01
P02
P03
P04
P05
P06
P07
P08
P09
P10
P11
P12
P13
P14
P15
IOP0
IOP1
FEP0
FEP1

【図20】

図 20



フロントページの続き

(72)発明者　濱中　直樹
東京都国分寺市東恋ケ窪１丁目280番地
株式会社日立製作所中央研究所内
(72)発明者　千葉　寛之
東京都国分寺市東恋ケ窪１丁目280番地
株式会社日立製作所中央研究所内
(72)発明者　樋口　達雄
東京都国分寺市東恋ケ窪１丁目280番地
株式会社日立製作所中央研究所内
(72)発明者　首藤　信一
東京都国分寺市東恋ケ窪１丁目280番地
株式会社日立製作所中央研究所内

(72)発明者　緒方　康洋
東京都小平市上水本町５丁目20番１号　日立超エル・エス・アイ・エンジニアリング株式会社内
(72)発明者　武内　茂雄
東京都小平市上水本町５丁目20番１号　日立超エル・エス・アイ・エンジニアリング株式会社内
(72)発明者　鳥羽　達
東京都小平市上水本町５丁目20番１号　日立超エル・エス・アイ・エンジニアリング株式会社内